Characterisation of the metabolome of *Mycobacterium tuberculosis* to identify new pathways and pathway holes

by

Kristen Marie Wolfenden

WLFKRI002

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN

In fulfillment of the requirements for the degree

MSc Bioinformatics

Faculty of Health Sciences

UNIVERSITY OF CAPE TOWN

Submitted 17 February 2014

Supervisor Nicola Mulder

Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town

DECLARATION

I, Kristen Wolfenden, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work, nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: ……..

Date:……..

Abstract

Background: Due to high incidence rates and the development of new drug-resistant or multidrug-resistant strains of TB, the development of new medicines and treatments for tuberculosis is a necessity. In order to develop these drugs, *Mycobacterium tuberculosis* (Mtb) needs to be studied more completely; this study performs a characterisation of the metabolome of Mtb and comparison across the phylogenetic profile to identify notable pathways.

Methods & Materials: To unravel their roles in the cell, data has been integrated from a variety of sources, both computationally predicted and experimental, to generate metabolic networks for the proteome. This first step involved creating a more complete catalogue of Mtb metabolic pathways and their phylogenetic profiles. The data from multiple resources was compiled together into a matrix of all Mtb strain H37Rv proteins, along with unique (non-orthologous) proteins from strains KZN 1435, CDC1551, H37Ra and F11. For each protein, we derived a full phylogenetic profile of over 1,000 organisms with all orthologs in the five Mtb strains. Pathway data was initially filled from KEGG and UniPathway data, and then additional proteins were annotated using Gene Ontology (GO) terms, KEGG reference pathways and homology to characterized proteins from closely related organisms. Following this, pathway information was compared between Mtb and *M. leprae*, *C. glutamicum, E. coli*, *H. sapiens* and all organisms across the profile to identify notable pathways for potential drug targets. Next these functional networks were used in an attempt to find pathway holes, which could be circumstances where Mtb utilises the host genome to accomplish its metabolic needs.

Results: A total of 553 Mtb proteins were added to previously existing pathways based on EC number and GO terms, and 288 reactions were characterized with enzymatic proteins based on sequence homology from BLASTP matches with closely related organisms. A number of interesting pathways were identified fulfilling these characteristics including arginine and proline metabolism and glycerolipid metabolism (many orthologs shared between Mtb and M. leprae), histidine metabolism and

butanoate metabolism (few orthologs shared between Mtb and E. coli), xylene degradation and porphyrin and chlorophyll metabolism (many orthologs shared between Mtb and C. glutamicum) and lastly biotin metabolism, sulfur metabolism, geraniol degradation and polycyclic aromatic hydrocarbon degradation (few orthologs between Mtb and H. sapiens)

Conclusion: This study contributed to increasing the number of characterised enzymes and pathways in Mtb and the results suggest that the functional annotation of Mtb needs to be updated within the KEGG database. Many proteins were functionally annotated and notable pathways described for further research. It is hoped that this more complete characterization and identification of potential drug targets can aid in the understanding of the metabolism of Mtb and lead to new drug development.

Contents

Abstract i
Contents iii
List of Figures vii
List of Tables viii
Acknowledgements ix
List of Abbreviations and Acronyms x
1 Introduction 1
1.1 Epidemiology 1
1.2 History 2
1.3 Mycobacteria 3
1.3.1 *Mycobacterium tuberculosis* Complex 4
1.3.2 Phylogeny and Related Organisms 6
1.4 Tuberculosis the Disease 7
1.4.1 Transmission 7
1.4.2 Infection 7
1.4.3 Diagnosis 8
1.5 Reducing the Prevalence and Incidence of Tuberculosis 9
1.5.1 Public Health Measures 9
1.5.2 Drugs 9
1.5.3 Vaccines 11
1.5.4 Other Treatment Options 11
1.6 Complications Encountered When Treating TB 11
1.6.1 Persistence 12
1.6.2 Slow Diagnosis 14
1.6.3 Patient Compliance 15
1.6.4 HIV Coinfection 15
1.6.5 Drug-Resistance 15
1.7 Bioinformatics 16
1.7.1 Genome Sequencing 16
1.7.2 Databases 18
1.7.3 Tools 21

1.8 Annotations 23
1.8.1 Automatic Annotation 23
1.8.2 Manual Annotation 24
1.9 Metabolic Pathways 24
1.9.1 Functional Classification 24
1.9.2 Comparison of Pathways 25
1.10 Pathway Holes 26
1.11 Need for New Drugs 26
1.12 Thesis Rationale 27
1.13 Aims of the Study 27
1.14 Road Map 28
2 Materials and Methods 29
2.1 Extraction of Ortholog Data 29
2.2 Incorporating Additional Orthologs 30
2.3 Removal of Orthologous Proteins in H37Ra, F11, KZN and CDC1551 31
2.4 Manual Annotation of Proteins without Pathway Data 31
2.4.1 UniProt Search 32
2.4.2 KEGG Search 32
2.4.3 KEGG BRITE Hierarchies 32
2.4.4 Converting UniPathway Pathways Into KEGG Pathways 33
2.4.5 Using EC Numbers to Find Additional Pathway Memberships 34
2.4.6 Using GO Terms and Gene Names to Derive Additional Pathways 35
2.4.7 Additional Pathways from UniPathway 35
2.5 Reordering of Matrix 36
2.6 Mapping Proteins to KEGG Pathway Diagrams 36
2.7 Deriving Additional Pathways Using BLASTP 36
2.7.1 Using BLAST on Other Organisms 36
2.8 Analysing the Phylogenetic Profile 37
2.9 Deriving the Pathway Summary Totals for *Escherichia coli* 38
2.10 Identifying 'Missing' Pathways or Pathway Holes 38
2.10.1 Pathway Tools 38
2.10.2 KEGG 38
2.10.3 Existence in Closely Related Organisms 38
2.10.4 Resemblance 39

2.11 Conclusion 39
3 Results 40
3.1 Extraction of Ortholog Data 40
3.2 Incorporating Additional Orthologs 40
3.3 Removal of Orthologous Proteins in H37Ra, F11, KZN and CDC1551 41
3.4 Manual Annotation of Proteins without Pathway Data 41
3.4.1 UniProt Search 42
3.4.2 KEGG Search 42
3.4.3 KEGG BRITE Hierarchies 42
3.4.4 Converting UniPathway Pathways Into KEGG Mapping Pathways 42
3.4.5 Using EC Numbers to Find Additional Pathway Memberships 43
3.4.6 Using GO Terms and Gene Names to Derive Additional Pathways 43
3.4.7 Additional Pathways from UniPathway 43
3.5 Reordering of Matrix 44
3.6 Mapping Proteins to KEGG Pathway Diagrams 44
3.7 Annotating Additional Reactions Using BLASTP 45
3.7.1 Using BLASTP on Other Organisms 45
3.8 Analysing the Phylogenetic Profile 50
3.9 Deriving the Pathway Summary Totals for *Escherichia coli* 53
3.10 Identifying 'Missing' Pathways or Pathway Holes 56
3.10.1 Pathway Tools 56
3.10.2 KEGG 57
3.10.3 Existence in Closely Related Organisms 57
3.10.4 Visual Evidence for Pathway Holes 57
3.11 Conclusion 57
4 Discussion 58
4.1 Assessment of Methods Used 58
4.2 Interesting Pathways Filled in *M. tuberculosis* 61
4.2.1 Benzoate Degradation 61
4.2.2 Phenylalanine Metabolism 69
4.2.3 Glyoxylate and Dicarboxylate Metabolism 74
4.3 Comparing Pathways between Organisms 80
4.3.1 Many Proteins Added to Both *M. tuberculosis* and *M. leprae* 82
4.3.2 Many Proteins Added to *M. tuberculosis* But Not *M. leprae* 84

4.3.3 Large Differences Between *M. tuberculosis* and *E. coli* 85
4.3.4 Many Similarities Between *M. tuberculosis* and *C. glutamicum* 87
4.3.5 Large Differences Between *M. tuberculosis* and *H. sapiens* 88
4.4 Missing Pathway Results 91
4.5 Conclusion 93
5 Conclusion 95
References 100
Appendix A 113
Appendix B 119
Appendix C 122

List of Figures

Figure 1.1 Tuberculosis around the world..... 2

Figure 1.2 Phylogeny of the order *Actinomycetales*. 6

Figure 1.3: This image shows all the commonly used drugs to treat tuberculosis infection, including their method of action (Zhang, 2005)..... 10

Figure 1.4 Pseudoorthology and xenoparalogy caused by horizontal gene transfer. 18

Figure 2.1 Example of UniPathway pathway membership information..... 33

Figure 2.2 Pathway membership tree view in UniPathway. 34

Figure 2.3 Example of a pathway classified as 'missing' or a pathway hole..... 39

Figure 3.1: Heat map of the frequencies of proteins per pathway in each organism..... 52

Figure 3.2 These charts show the average frequency of homologs for all species in the phylogenetic profile according to pathway..... 53

Figure 3.3 Bar graphs (A-F) comparing the numbers of proteins per pathway for KEGG reference pathways, *M. tuberculosis* H37Rv, *M. leprae* TN, *C. glutamicum*, *E. coli* and *H. sapiens*. 55

Figure 4.1 This image shows the general aerobic metabolism of aromatic compounds.. 62

Figure 4.2 The KEGG pathway for *M. tuberculosis* H37Rv benzoate degradation……65

Figure 4.3 Figures showing the chromosome location of four newly annotated Mtb H37Rv genes to two reactions that have one reaction in between. 68

Figure 4.4 The KEGG pathway for M. tuberculosis H37Rv phenylalanine metabolism…70

Figure 4.5 These images show the chromosome locations for genes encoding four proteins in Mtb H37Rv. 73

Figure 4.6 The KEGG pathway for M. tuberculosis H37Rv glyoxylate and dicarboxylate metabolism……75

Figure 4.7 Image showing the location of two genes encoding proteins in the 'glyoxylate and dicarboxylate metabolism' pathway. 80

List of Tables

Table 2.1 Example structure of the compiled matrix. 29

Table 3.1 This shows the number of orthologs for the five strains of *M. tuberculosis* in the original and updated phylogenetic profile. 41

Table 3.2 Table showing the number of unique proteins from each strain in the phylogenetic profile. 41

Table 3.3 Table showing the numbers of proteins annotated by each manual annotation method. 43

Table 3.4 Results of annotations for proteins of *M. tuberculosis* H37Rv. 47

Acknowledgements

This work was carried out at the Computational Biology Group (CBIO) in the Health Sciences Faculty at the University of Cape Town. I would like to give a huge thanks to my supervisor, Nicola Mulder, who kept helping and supporting me even when I disappeared for periods of time. Without her guidance I would have been lost, and I sincerely thank her for her patience and for never giving up on me.

To Olivier, my fiancé and soon to be husband, I would never have finished without you. If you had not been there to push me, to encourage me, to lift me up when I felt all was lost or even to cook for me, this project would not be what it is today. You cut through all my highly developed procrastination techniques and pushed me to the finish line. I thank the Lord every day for having brought you into my life and know that we will share our love and hope for the rest of our lives.

To my family, Dad, Mom, Elisabeth, Charlie, Julie and Bobby, whose incessant 'are you finished yet?'s stressed me out but also thrusted me forward. Thank you for supporting me, for giving me all the love a girl could ever ask for, and stressing me when I needed motivation.

To my good friends Thembi Dube and Kamo Direko, for pushing me on, giving me advice and always being there to help me see through the tough times. Also all of my basketball buddies, such as the girls of the UCT Basketball Team and Lethal Ladies; thank you for keeping me sane and giving me an outlet! The last several years have not been easy, but with you all in my life I have been able to weather through and finish this thesis.

Last but not least, to all the lab friends I have made, and especially Cashifa for always being so helpful whenever it comes to administrative matters or whenever I needed a 'girl talk'. Thank you all for providing such a stimulating environment!

Kris Wolfenden ☺

List of Abbreviations and Acronyms

BCG Bacillus Calmette-Guérin

BLAST(P) Basic Local Alignment Search Tool (Protein)

CS Cycloserine

DNA Deoxyribonucleic Acid

DOTS Directly Observed Treatment, Short-Course

EC Enzyme Commission

EMB Ethambutol

EMBL European Molecular Biology Laboratory

ENA European Nucleotide Archive

ETH Ethionamide

FQ Fluoroquinolone

GO Gene Ontology

HGT Horizontal Gene Transfer

HIV Human Immunodeficiency Virus

ICL Isocitrate Lyase

IGRA Interferon Gamma Release Assay

INH Isoniazid

KEGG Kyoto Encyclopaedia of Genes and Genomes

KO KEGG Orthology

MDR Multidrug-Resistant

MeV Multi-Experiment Viewer

Mtb *Mycobacterium tuberculosis*

MTBC Mycobacterium tuberculosis Complex

NCBI-GI National Centre for Biotechnology Information- GenInfo Identifier

PAS Para-Aminosalicylic Acid

PGDB Pathway/Genome Database

PPD Purified Protein Derivative

PT Pathway Tools

PTH Prothionamide

PZA Pyrazinamide

RIF Rifampin

RNA Ribonucleic Acid

SM Streptomycin

TB Tuberculosis

TST Tuberculin Skin Test

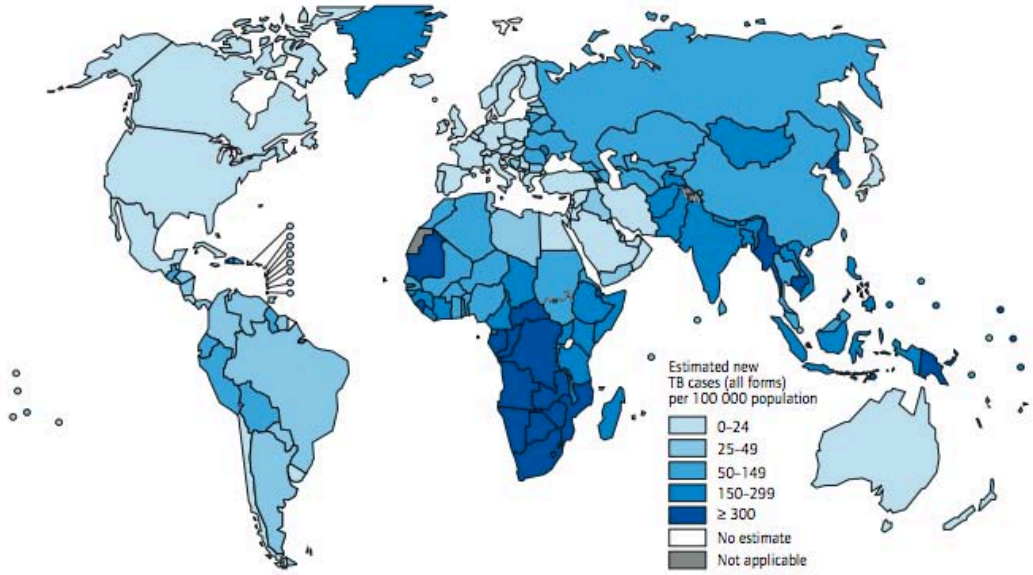
WHO World Health Organization

XDR Extremely Drug-Resistant

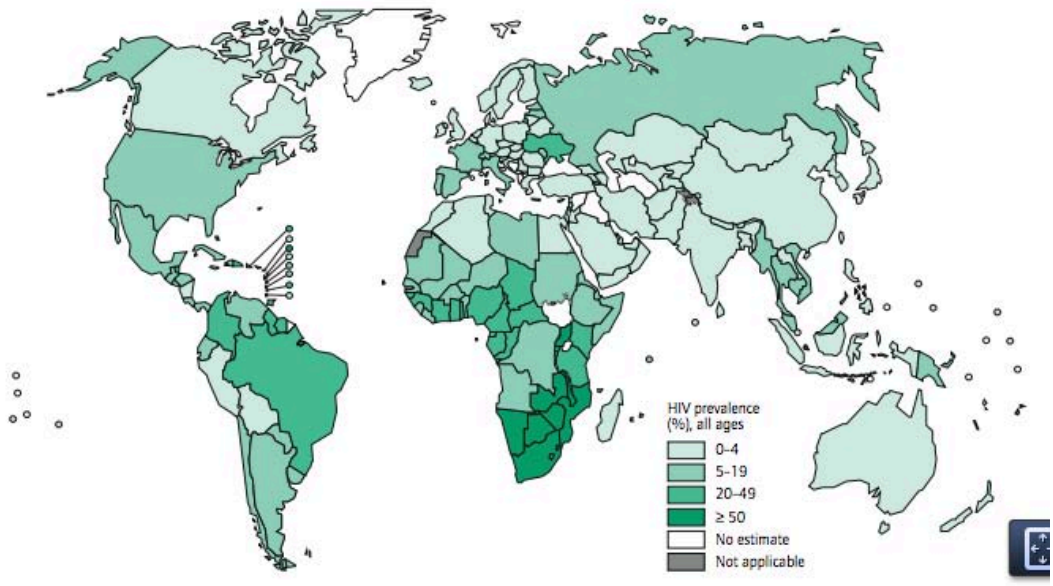
1 Introduction

1.1 Epidemiology

Tuberculosis (TB), an airborne contagious disease caused by *Mycobacterium tuberculosis (*Mtb*)*, is an ancient disease that has probably infected humans for millions of years (Gutierrez et al., 2005). As well as being ancient, TB infects humans worldwide, killing 1.4 million people per year and causing 8.7 million new cases in 2011 alone (World Health Organization 2012). It is also a disease of poverty with 95% of deaths occurring in the developing world, affecting mostly young adults at the most productive time of their lives (World Health Organization 2012a). Additionally, resistance in Mtb occurs regularly; the World Health Organization (WHO) estimates that there were half a million new multidrug-resistant cases in 2011 (World Health Organization 2013). Lastly, co-infection with both TB and HIV is common and causes additional problems when diagnosing and treating TB; in fact one third of the 34 million HIV-positive people in the world are co-infected with latent TB and one in four HIV-related deaths is due to TB (World Health Organization 2013a). In South Africa, 65% of TB patients are also HIV positive, making the country an epicentre of the co-epidemic (World Health Organization 2013c). Below are two maps which show first the incidence of TB per 100,000 population, and secondly the estimated HIV prevalence among TB cases.



A.



B.

Figure 1.1 Tuberculosis around the world. The first image (A) displays the incidence of tuberculosis while the second (B) shows HIV prevalence amongst cases of TB reference (World Health Organization, 2012a, 2012b).

These maps show both the high incidence of TB and the high prevalence of HIV among new cases of TB in sub-Saharan Africa and South Africa in particular (World Health Organization 2012b). In order to reduce this high incidence and prevalence shown, further studies need to be performed on the causative agent, *Mycobacterium tuberculosis.*

1.2 History

The genus *Mycobacterium* is believed to have originated as many as 150 million years ago (Daniel, 2006). This genus includes many different species including fast and slow-growing mycobacteria and nontuberculous and tuberculous mycobacteria. An ancestor of *Mycobacterium tuberculosis* was present as early as 3 million years ago in East Africa, suggesting that it might have affected early hominids (Gutierrez et al., 2005). Today it is believed that all modern members of the Mtb complex (MTBC), which includes *Mycobacterium bovis, Mycobacterium canettii* and *Mycobacterium africanum* as well as Mtb, had a common ancestor in Africa 35,000-15,000 years ago. Modern strains of Mtb are thought to have evolved from a common ancestor around 20,000-15,000 years ago (Daniel, 2006). The earliest evidence of the disease is in human remains from around 5000 BC, which show deformities in the spinal column indicative of extrapulmonary tuberculosis (Kaufmann and van Helden, 2008). During Greek times, around 460AD,

Hippocrates wrote that phthisis, an older name for tuberculosis, was the most widespread disease at the time (Daniel, 2006). Three years after Louis Pasteur developed the germ theory of infectious disease in 1862, Jean-Antoine Villemin experimentally proved that tuberculosis could be inoculated from man or cow to laboratory animals such as rabbits and guinea pigs. Villemin also showed that sputum from an infected patient could infect a rabbit with tuberculosis (Sakula, 1983). The turning point in this battle against TB came on 24 March 1882, when Robert Koch identified and described the bacteria causing TB, *Mycobacterium tuberculosis,* at a meeting of the Physiological Society of Berlin (Koch et al., 1982). He described the bacteria as being "rod-shaped" and "very thin", and with a length of usually "one-fourth to one-half as long as the diameter of a red blood cell"; he found these bacilli "ordinarily form small groups of cells which are pressed together and arranged in bundles" and remarked upon their similarity to the bacilli of leprosy (caused by *Mycobacterium leprae*) (Koch et al., 1982). By the 1930s Florence Seibert developed purified protein derivative (PPD), used to test for the presence of tuberculosis; PPD was then used by the World Health Organization in 1955 to demonstrate the existence of latent tuberculosis infections, first noted in 1909 by Clemens Freiherr von Pirquet, in healthy school children in countries with a high prevalence of tuberculosis. Even with this long history of infection in humans, the first effective treatment for the disease was only developed in 1944 with the isolation of streptomycin; isoniazid and rifamycins, also effective treatment drugs for tuberculosis, were developed in 1952 and 1957, respectively (Daniel, 2006). Since the 1950s, however, few new effective drugs have been produced, and work must continue to develop new chemotherapies for the treatment of tuberculosis and especially drug-resistant tuberculosis (Kaufmann and van Helden, 2008).

1.3 Mycobacteria

The genus *Mycobacterium* includes more than 140 described species, all of which show at least 94.3% sequence similarity in their 16S ribosomal RNA genes. Mycobacteria are aerobic, acid-fast actinomycetes and usually take the form of straight or slightly curved rods; they have waxy and hydrophobic cell walls that are thicker than most other bacteria as well as being rich in mycolic acids. They can be categorized into fast and slow-growing mycobacteria; fast-growing mycobacteria are typically avirulent and

include *Mycobacterium abscessus* and *Mycobacterium smegmatis* while slow growing cause many human and animal diseases and include Mtb and *Mycobacterium leprae* (Hartmans et al., 2006). Current evidence shows that slow growing mycobacteria evolved from a single ancestral fast-growing species (Tortoli, 2012). One phylogenetic branch within the slow-growers incorporates the smooth tubercle bacilli, named smooth because they form smooth colonies, and the *Mycobacterium tuberculosis* complex (MTBC) bacteria; bacteria in the MTBC are the most common causes of tuberculosis in humans although smooth strains have also been isolated from human cases. Most smooth species have been isolated from locations in East Africa and it has been estimated that to accumulate the observed level of synonymous nucleotide variation within tubercle bacilli would have taken at least 2.6 to 2.8 million years to develop; this suggests that the evolution of tubercle bacilli may have taken place in the same location and over the same time period as the evolution of humans, and moved out of East Africa along with the migrations of humans. In addition, while smooth species show a high ratio of synonymous to nonsynonymous substitution showing purifying selection over time against amino acid changes, species of the MTBC have a very low ratio suggesting recent expansion. Thus, strains of the MTBC may have recently evolved from an ancestral-type similar to the smooth species of tuberculosis (Gutierrez et al., 2005).

1.3.1 *Mycobacterium tuberculosis* Complex

The MTBC includes the Mtb*, M. bovis, M. microti, M. africanum, M. pinnipedii* and *M. caprae* species. The MTBC is a clonal group that shows no recombination, unlike smooth strains of tubercle bacilli (Garcia-Betancur et al., 2012). They appear to have originated from a number of horizontal gene transfer events and then went through an evolutionary bottleneck and finally clonal expansion about 35,000 years ago (Gutierrez et al., 2005). Species of the MTBC show an extremely low level of genetic variation, with more than 99.95% sequence similarity at the nucleotide level (Smith et al., 2009). Some would consider bacteria of the MTBC as separate species while other reports show that these species may in fact belong to only one genospecies, owing to the identical nucleotide sequences of the 16SrRNA and rpoB genes, usually used to differentiate species in mycobacteria. Additionally, of the MTBC members, studies of both genomic signatures and multiple whole-genome alignments show that these bacteria could

belong to a single species. Of the MTBC, only KZN 1435 showed a single central inversion (Garcia-Betancur et al., 2012). Different species within the MTBC can be differentiated according to their distinct host preference. For example, Mtb and *M. africanum* subtype 1 are human adapted, while *M. bovis* is mainly found in cattle and *M. pinnipedii* is found in marine mammals. However, this host preference does not always hold true; all species have been isolated from cases of human tuberculosis and many species have been isolated from mammals not identified as the primary host (Smith et al., 2009). It is most likely that animal-adapted species of the MTBC developed from a human-adapted ancestral strain of Mtb. Thus, the disease would have originated in humans and subsequently evolved to infect other species (Smith et al., 2009).

The biggest differences between Mtb*, M. leprae, M. avium*, and *M. bovis* are in the cell wall products and the PE/PPE/PGRS proteins, but there are also differences in lipid metabolism, cell wall proteins, insertion sequence elements and hypothetical proteins. *M. tuberculosis* H37Rv has about 3,900 genes encoding proteins while *M. leprae*, with a substantially reduced genome, has about 1,650. *M. leprae* also shows many genomic rearrangements in comparison to Mtb. *M. tuberculosis* H37Rv was shown to have two unique genes compared to the other strains of *Mycobacterium*, while *M. tuberculosis* CDC1551 has 122 and *M. leprae* has 149. All genomes have a functional glycolytic pathway and tricarboxylic acid (TCA) cycle. Genes involved in folic acid, pantothenate, pyridoxine, thiamine biosynthesis and purine and pyrimidine biosynthesis are also highly conserved. *M. leprae* has lost many of the genes involved in lipid metabolism, but mycolic acid biosynthesis is highly conserved; *M. leprae* has also lost nitrate and nitrite reductase, fumarate reductase, and the urease and NADH oxidase operons, limiting growth under anaerobic and microaerophilic conditions. Lipid metabolism and cell wall proteins could be related to virulence as many lipids function in the cell membrane, the interface between host and pathogen. Additionally, polyketide synthases show wide differences among the organisms, with *M. leprae* having lost many of the genes. Mtb has many PE and PPE genes, constituting 10% of its genome, while others have lower numbers of these genes and *M. leprae* has very few. Lastly, there are relatively high intra-species differences in Mtb; this could also lead to differences in pathogenesis (Marri et al., 2006).

1.3.2 Phylogeny and Related Organisms

Mycobacteria belong to the phylum *Actinobacteria*, order *Actinomycetales*, suborder *Corynebacterineae,* family *Mycobacteriaceae* (Zhi et al., 2009). Order *Actinomycetales* contains a diverse group of soil-growing, marine, plant sybiont and parasitic bacteria (Alam et al., 2010).

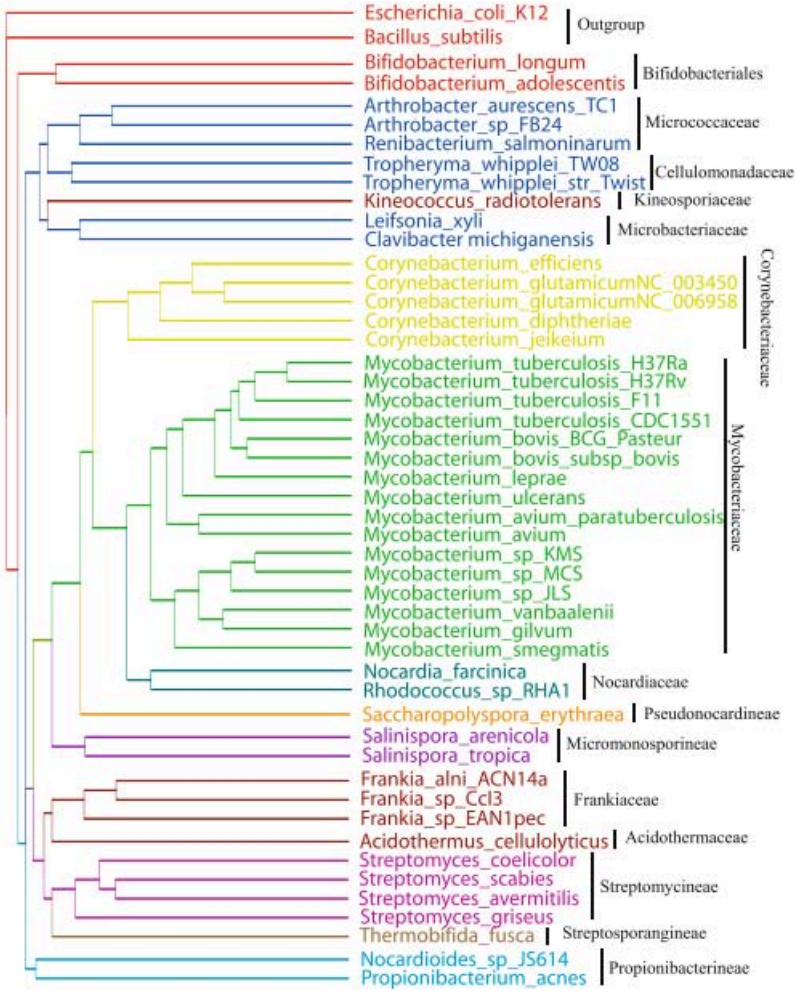


Figure 1.2 Phylogeny of the order *Actinomycetales*. This shows the major groups within the order *Actinomycetales* including the various families, and species of those families (Alam et al., 2010).

A number of organisms can be seen on the phylogeny above, including the closely related organisms *Corynebacterium glutamicum* (a soil actinomycete extremely important in industrial production of amino acids) (Kalinowski et al., 2003), *Nocardia farcinica* (an infectious aerobic actinomycete causing nocardiosis in humans) (Torres et al., 2000), *Rhodococcus* sp. (including *Rhodococcus jostii*, a soil actinomycete useful in bioremediation, and *Rhodococcus erythropolis,* a soil actinomycete known for its ability to degrade alkanes) (LeBlanc et al., 2008; Sekine et al., 2006) and *Streptomyces*

coelicolor (important in the biotech industry for the production of antibiotics) (Alam et al., 2010). The genera *Gordonia,* which contains *Gordonia bronchialis,* also belongs to the family *Nocardiaceae* (Zhi et al., 2009).

1.4 Tuberculosis the Disease

For tuberculosis to infect a human and be diagnosed multiple stages must take place. First the disease must be transmitted; secondly an infection develops, and lastly diagnosis must occur.

1.4.1 Transmission

Tuberculosis is primarily a respiratory disease spread by inhaling droplets containing the bacterium that have been coughed or sneezed out by a person infected with infectious tuberculosis. Infectious tuberculosis means the individuals have 'smear-positive' sputum, or sputum in which bacteria can be easily seen with a microscope (National Institute for Health and Clinical Excellence, 2011).

1.4.2 Infection

Once the bacterium enters the lungs, it slowly begins to grow. At this stage, the immune system usually responds; for 80% of infected people the immune system successfully kills the bacteria and removes them from the body. In some cases the immune system is unable to kill the bacteria but successfully creates a surrounding barrier to prevent the bacteria from growing, and the bacteria enter a dormant state (National Institute for Health and Clinical Excellence, 2011). Once this happens, a person has latent TB infection, which can last for months, years or even decades. Someone with latent TB infection may never develop the disease nor show any symptoms but may still need treatment in order to prevent the disease from becoming active at a later point in time (US National Institute of Allergy and Infectious Disease, 2009). Only about one in ten individuals infected with TB will develop active tuberculosis in their lifetime; this occurs when the immune system fails to contain the bacteria or the barriers break down at a later point in time. Oftentimes active TB strikes those with weakened immune systems, such as HIV infected individuals or the very young or elderly. Active TB typically involves infection of the lungs and will present as lesions on the lungs, coughing, fever and/or weight loss; this is called pulmonary TB. Extrapulmonary TB occurs when bacteria are carried by the blood to other tissues in the body, and can

affect all other bodily tissues including bones, the spine, kidneys, lymph nodes and the brain; symptoms include pleural effusion, meningitis, miliary disease and pericardial effusion (Dye et al., 2006).

1.4.2.1 Macrophage Environment

During infection, many bacteria exist intracellularly within early phagosomes of macrophages by arresting development of the phagosome (Kaufmann and Parida, 2008). The environment of the macrophage early phagosome is subject to debate but it is believed that this environment is characterised by hypoxia, relatively high pH of 6.2 and nutrient scarcity restricting the bacteria to the use of fatty acids as a source of carbon (Galagan et al., 2013; Guirado et al., 2013). Indeed, Mtb preferentially metabolises lipid substrates over carbohydrates *ex vivo* so it probably has distinctive central carbon metabolism that helps it adapt to the macrophage environment (Rhee et al., 2011). Additionally, Mtb has been shown to hijack the host metabolism to acquire some essential nutrients and prevent its degradation by host mechanisms (Pieters, 2008).

1.4.3 Diagnosis

TB diagnosis in patients can occur through a number of diagnostic tests. The first is the tuberculin skin test (TST) in which a tiny amount of Mtb protein is injected under the skin; if there is a reaction, such as if the spot turns into a raised red mark, then the person has been exposed to the bacterium. However, this only shows exposure to the bacteria and not whether or not the person has active TB. Chest radiography is often used to determine if a person has the active disease because lesions or other signs would be observed on the x-rays. Sputum smear specimens are the final most common and inexpensive diagnosis method in which sputum samples are taken from the patient and examined under the microscope. Sputum smear microscopy becomes ineffective, however, if the number of bacteria in the sputum is low (Dye et al., 2006). All these diagnosis methods are older methods, and because of their low sensitivities, new methods have been developed. Modern diagnosis methods include tissue biopsies, cultures and ultrasounds, as well as molecular methods such as nucleic acid amplification (Xpert MTB/RIF system) (World Health Organization, 2013d). Cultures

are more typically used for drug resistance testing rather than diagnosis as the slow-growing bacteria can take two to three months to grow in a culture.

1.5 Reducing the Prevalence and Incidence of Tuberculosis

Tuberculosis is an entirely treatable disease when the correct drugs are administered to patients; with correct treatment there is a cure rate of 90% for patients with drug-susceptible strains and 50-70% for patients with multidrug-resistant strains (World Health Organization, 2011). However, treatment is complicated by factors such as the length of time drugs need to be taken (up to two years), drug-resistance, immunosuppressive diseases such as HIV and reactivations of latent tuberculosis. Thus a number of strategies have been implemented in order to combat the spread of TB around the world.

1.5.1 Public Health Measures

Tuberculosis has a greater effect on the poor throughout the world, explained by lower levels of access to health care, poor living conditions, malnutrition, and HIV infection. Other risk factors include smoking, alcohol abuse, diabetes mellitus and drug abuse. Improving socioeconomic conditions helps to control the spread of TB (World Health Organization, 2011). In addition, governments and international groups need to ensure that tuberculosis control programs are adequately funded, health care providers are appropriately trained and supported, infection control practices are implemented, second-line drugs are made available, communities facing certain risk factors have programs tailored to their particular needs, and that they engage communities in controlling tuberculosis (Abubakar et al., 2013).

1.5.2 Drugs

Most of the drugs used today to treat tuberculosis were originally developed in the 1940s and 1950s. The first effective drug against tuberculosis, streptomycin (SM), was developed in 1944 (Schatz, 2005). Two years later, para-aminosalicylic acid (PAS) was discovered (Lehmann, 1946). Isoniazid (INH), a highly active, inexpensive antituberculosis drug with minimal side effects, was developed in 1952. Pyrazolinamide (PZA) came shortly after in 1952, with ethionamide (ETH)/prothionamide (PTH) following in 1956. Ethambutol (EMB) was then discovered in 1961, followed by many other drugs including cycloserine (CS), kanamycin and amikacin, viomycin,

capreomycin and rifamycins, developed into rifampin (RIF) the most common used drug to treat tuberculosis since the 1970s. Quinolone drugs (FQ) were only developed in the 1980s based on research conducted in the 1960s. Of these drugs, INH, RIF, PZA, EMB and SM are considered first-line drugs, whereas kanamycin, amikacin, capreomycin, CS, ETH/PTH, thiacetazone and FQ are second-line drugs. Lastly, some of these drugs are specific for TB or mycobacteria such as INH, PZA, PAS, ETH, EMB and thiacetazone, while the rest are broad-spectrum antibiotics (Zhang, 2005).

Drug (year of discovery)	MIC[a] (g/ml)	Effect on bacterial cell	Mechanisms of action	Targets	Genes involved in resistance
Isoniazid (1952)	0.01–0.2	Bactericidal	Inhibition of cell wall mycolic acid synthesis and other multiple effects on DNA, lipids, carbohydrates, and NAD metabolism	Multiple targets including acyl carrier protein reductase (InhA)	*katG*[b] *inhA* *ndh*
Rifampin (1966)	0.05–0.5	Bactericidal	Inhibition of RNA synthesis	RNA polymerase β subunit	*rpoB*
Pyrazinamide (1952)	20–100 pH 5.5 or 6.0	Bacteriostatic/ bactericidal	Disruption of membrane transport and energy depletion	Membrane energy metabolism	*pncA*[b]
Ethambutol (1961)	1–5	Bacteriostatic	Inhibition of cell wall arabinogalactan synthesis	Arabinosyl transferase	*embCAB*
Streptomycin (1944)	2–8	Bactericidal	Inhibition of protein synthesis	Ribosomal S12 protein and16S rRNA	*rpsL*, *rrs*
Kanamycin (1957)	1–8	Bactericidal	Inhibition of protein synthesis	16S rRNA	*rrs*
Quinolones (1963)	0.2–4	Bactericidal	Inhibition of DNA synthesis	DNA gyrase	*gyrA* *gyrB*
Ethionamide (1956)	0.6–2.5	Bacteriostatic	Inhibition of mycolic acid synthesis	Acyl carrier protein reductase (InhA)	*inhA* *etaA/ethA*[b]
PAS (1946)	1–8	Bacteriostatic	Inhibition of folic acid and iron metabolism?	Unknown	Unknown
Cycloserine (1952)	5–20	Bacteriostatic	Inhibition of peptidoglycan synthesis	D-alanine racemase[c]	*alrA*, *Ddl*[c]

[a]MIC is based on Inderlied & Salfinger (13).
[b]KatG, PncA, and EtaA/EthA are enzymes involved in the activation of prodrugs INH, PZA, and ETH, respectively.
[c]In fast growing *M. smegmatis*.

Figure 1.3: This image shows all the commonly used drugs to treat tuberculosis infection, including their method of action (Zhang, 2005).

Due to complications arising from drug-resistance and the life history of Mtb, (that it can form persistent populations of bacteria that can lie dormant for many years in hosts until being reactivated) TB treatment typically involves a combination of drugs given over a long period of time. The most standard form of therapy is the Directly Observed Treatment, Short-course (DOTS), developed by the World Health Organization (WHO) for treatment of drug-sensitive tuberculosis. DOTS involves a six month regimen including four first-line drugs. During the first two months, INH and RIF, the two most powerful drugs against TB, plus EMB and PZA are all taken, and then the last four months INH and RIF are continued. DOTS typically cures around 90% of drug-

susceptible TB cases; however, many cases now are characterised by drug-resistance, leading to more complicated and longer treatments. Treatment of drug resistant cases involves both first-line and second-line drugs and can take up to two years for a patient to complete treatment (World Health Organization, 2011).

1.5.3 Vaccines

Currently the only vaccine in regular use for preventing tuberculosis is the bacillus Calmette-Guérin (BCG) vaccine, which has been in use since 1921 (Colditz et al., 1994). However, this vaccine typically provides protection only against childhood forms of TB, particularly meningitis and military disease, but minimal or no protection against pulmonary TB in adults (Trunz et al., 2006). Even with high levels of BCG vaccination, there still exists a high incidence of tuberculosis in children in endemic countries like South Africa (Tameris et al., 2013). Additionally, BCG vaccine can harm HIV positive infants and should not be administered to these children. A number of vaccines are being developed which are either designed to replace BCG as a recombinant live vaccine or act as a booster for BCG, but these are likely to have limited capabilities at preventing infection and preventing activation post-exposure (Kaufmann et al., 2010). Additionally, immunotherapeutic vaccines that can boost the immune system during treatment are currently being tested (Jassal and Bishai, 2009).

1.5.4 Other Treatment Options

Treatment of tuberculosis could also include surgery in some cases. Surgical intervention might be considered if the patient has significant localised pulmonary tuberculosis. This option would likely be performed in infections with high levels of resistance to drugs (Jassal and Bishai, 2009).

1.6 Complications Encountered When Treating TB

As a bacterium that has evolved alongside humans for thousands of years, TB has proven extremely difficult to treat. Over the years, multiple discoveries have led to the belief that tuberculosis would soon be eradicated; for example, when BCG was first discovered it was thought to be the vaccine to prevent all tuberculosis. Similarly, after the discovery of streptomycin in the 1940s it was believed that TB treatment would thereafter be simple (Brennan and Thole, 2012). However, Mtb has proven highly resistant to eradication attempts, due to certain characteristics of both the bacteria and

management of the disease. Firstly, its ability to lie dormant for years in the host allows it to evade most current antituberculous drugs, which mainly attack the bacteria while growing, and then reactivate at a later stage when drugs are no longer present in the system (Zhang, 2005). Secondly, delays in diagnosis of the disease due to failing health care systems or poor diagnostic methods can increase transmission of the disease. Next, due to the long treatment duration, lack of drugs and sometimes serious side effects of drugs, patient compliance to treatment regimens is often poor, allowing further transmission and drug resistance to arise (Jassal and Bishai, 2009). In addition, the HIV epidemic requires new treatment plans so that drugs do not interact adversely, and so that new forms of extensively drug-resistant strains (XDR) do not develop in the absence of a strong host immune system. Lastly, Mtb has proven highly adaptable, evolving resistance to currently available drugs relatively easily, especially when there is slow diagnosis, poor adherence to treatment and HIV coinfection (Zhang, 2005).

1.6.1 Persistence

It is estimated that one third of the world is latently infected with Mtb; these individuals serve as a reservoir of the disease from which active infection can arise (Ma et al., 2010). It has been shown that lesions filled with tubercle bacilli include at least four different sub-populations. The first are actively growing bacteria, which are killed by INH, RIF or SM. Then there are bacteria with spurts of metabolism, killed by RIF, those in an acidic environment with low metabolic activity, killed by PZA, and finally those that are dormant and not killed by any currently used drug (Zhang, 2005). Dormant bacteria are the cause of such long periods of treatment. Bacteria in this dormant phase, also called persisters, are the cause of latent infection of tuberculosis, a state from which activation or reactivation of virulent TB can occur. Patients with latent infection have a 2-23% chance of reactivation of the disease in their lifetime (Zahrt, 2003). Latent infection can be diagnosed with a positive tuberculin skin test (TST) or positive result from another diagnostic tool (such as interferon gamma release assay (IGRA)) but no symptoms of the disease. These persisters typically exist in anaerobic conditions, and it has been thought that bacteria in this state are either slowly growing or non-replicating. Along with experiencing hypoxia, bacteria transitioning into the dormant state change their carbon source from glucose to fatty acids (Young et al., 2009). By existing intracellularly within early phagosomes of macrophages, they can cause a granuloma to

be formed around the infected macrophage (Kaufmann and Parida, 2008). This granuloma effectively walls off the complex from cytokines which can activate the macrophage, allowing the bacteria to persist for long periods of time (Bentrup and Russell, 2001).

A number of studies have attempted to elucidate the genes involved in the transition to and maintenance of dormancy, but this is still an on-going area of research. In reality it seems that persistence in Mtb requires the "coordinated expression of numerous virulence determinants, including those involved in intermediary and secondary metabolism, cell wall process, stress responses and signal transduction pathways" (Wang et al., 2011). Determining which genes are involved in this coordinated expression, however, remains difficult because there are few clinical models that can provide reliable data regarding mycobacterial metabolism during persistence. For example, *in vitro* models might not reflect the same conditions (energy source, presence of antibiotics, interactions with the host) experienced *in vivo* and thus gene expression patterns will be different (Dhar and McKinney, 2010). A recent study has refuted the idea that persisters come from subpopulations of nonreplicating bacteria that are unaffected by antibiotics. This study showed that there was no subpopulation of persisters prior to introduction of isoniazid, and rather that single-cell growth depended on cell size; it also showed that rather than having a semi-constant population of bacteria that survive through isoniazid treatment, the persistent subpopulation is in fact dynamic, and the population simply appears stable due to a balance between the rates of cell division and death (Wakamoto et al., 2013). Thus, persistent bacteria perhaps do not cease to replicate but rather form a stable population size by balancing cell division and death rates. Research regarding persistent populations of bacteria both in patients who have never had disease and have recovered from active disease continues to be performed, and will hopefully solve some of the questions and lead to better TB treatment methods.

In some cases latent infection may need to be treated with preventative therapy in order to avoid the development of active tuberculosis. In this case, isoniazid is administered for a period of 6-9 months. Preventative treatment can be particularly

effective in high-risk populations such as those who have had recent contact with someone who has active TB or for those who are HIV positive (Ma et al., 2010).

1.6.2 Slow Diagnosis

Within communities, the most effective strategy to reduce transmission of tuberculosis is early detection followed by quick and appropriate treatment (Abubakar et al., 2013). In particular, delays in diagnosis of active tuberculosis (latent tuberculosis is not thought to be infectious) can lead to a significant increase in transmission. The WHO estimates that a person with active tuberculosis can infect 10-15 other people per year through contact (World Health Organization, 2013e). With an increase in the reservoir of disease in the population, the bacteria have more opportunities to develop resistance to current drugs, which in turn makes treating the disease more difficult, ultimately furthering the spread of tuberculosis in a vicious cycle.

The most commonly used diagnostic method in most high-burden countries is sputum smear microscopy, which has low sensitivity, often missing cases in those with HIV coinfection, extrapulmonary dissemination, those with latent infection and children (Jassal and Bishai, 2009). For example, HIV-positive TB patients have been found to be smear-negative in 24%-61% of cases (Getahun et al., 2007). Newer diagnostic tests such as nucleic acid amplification and interferon gamma release assays, while providing more accurate and quicker results, are often not available or not used in many endemic areas due to high cost and limited capacities (Jassal and Bishai, 2009). Diagnostic tests to assess drug resistance usually include cultures, which can take 4-8 weeks to receive results. This delay increases the time allowed for transmission and the development of further drug-resistance by delaying the time spent until the appropriate treatment is administered (Jassal and Bishai, 2009). In fact, the WHO estimates that in 2011 only 19% of MDR-TB cases were found among notified cases of TB (World Health Organization, 2012a). While molecular methods to test for drug resistance are available, these diagnostic tests are typically more expensive and labour-intensive (Jassal and Bishai, 2009). However, the Xpert MTB/RIF molecular test has been adopted by some countries in order to diagnose TB and rifampicin resistance and should become more widespread over time (World Health Organization, 2012a). Better diagnostic methods that can diagnose both active and latent infection as well as drug resistance, increased

adoption of diagnostic methods, and improved infection control methods can disrupt transmission chains and significantly reduce the worldwide burden of tuberculosis.

1.6.3 Patient Compliance

Treatment for tuberculosis typically involves long time periods over which drugs are administered (at least six months). Additionally, some of the drugs, and particularly those administered for drug-resistant cases, have side effects that can discourage patients from adhering to the appropriate therapy. Long treatment period and negative side effects can result in irregular dosing and incomplete treatment (not completing the full six months of chemotherapy) which ultimately result in relapse and/or the development of drug resistance (Ma et al., 2010).

1.6.4 HIV Coinfection

HIV positive patients have both an increased likelihood of the disease progressing to active tuberculosis after infection and an increased likelihood of reactivation of latent infection (Jassal and Bishai, 2009). This is compounded by the fact that rifampicin, one of the most used first-line drugs to treat tuberculosis, interferes with many antiretroviral drugs (Ma et al., 2010). Thus, HIV coinfection further reduces TB treatment effectiveness.

1.6.5 Drug-Resistance

Drug susceptibility in mycobacteria is thought to be an acquired trait based on the environmental niche of the species. Since many of the antimycobacterial drugs are products of soil dwelling microbes, drug resistance is an adaptation to the environment in nontuberculous mycobacteria. Common human pathogenic mycobacteria do not usually come into contact with these microbes within the human body, and thus have adapted drug susceptibility, possibly increasing virulence. The most drug susceptible and virulent of mycobacteria include the MTBC, *M. kansasii*, *M. szulgai*, *M. marinum*, *M. malmoense, M. xenopi* and *M. leprae* (van Ingen et al., 2012). Even so, the emergence of drug resistance occurred almost immediately after the first drug, streptomycin, began being used to treat tuberculosis patients (Jassal and Bishai, 2009). After additional drugs were discovered, treatment regimens were formulated that seemed to prevent resistance from developing. However, by the 1980s new forms of multidrug-resistant strains began to develop; this emergence typically is caused by system failures such as

incomplete or inadequate treatment, lack of drugs leading to monotherapy, bad diagnostics and lack of regulation in accessing antibiotics. While these factors often result in acquired resistance, leading to reactivation within previously treated patients, many new cases of drug-resistance arise due to transmission of drug-resistant strains, particularly in areas of high incidence (Abubakar et al., 2013). Resistance can also be amplified, whereby a currently resistant strain develops increased levels of resistance due to inappropriate administration of drugs (Jassal and Bishai, 2009).

Additionally, on-going division during the latent phase may increase the likelihood of evolution of resistant strains. Drug resistance is not only caused by mutations but can also be caused by reversible phenotypic tolerance of drugs. It is possible that some of this tolerance is caused by epigenetic effects, as sister cells had a correlated survival rate under INH treatment (Jassal and Bishai, 2009).

Standardised treatment methods may not be appropriate for drug-resistant cases. This is certainly the case with XDR tuberculosis, where individualised treatment methods are necessary both in order to accurately treat the patient and to prevent further resistance from developing (Jassal and Bishai, 2009).

1.7 Bioinformatics

The publication of the human genome in February 2001 initiated a new era in the field of biology. Following this publication, a number of additional genomes have been published across a range of bacterial, plant and animal species. The combination of computer science, mathematics and statistics with biology has revolutionised the field and opened many opportunities for the advancement of genetics and disease research (Baxevanis and Ouellette, 2005). Biological research in bioinformatics is typically composed of three main aspects, databases, tools and algorithms. Various databases and tools have been used to complete this study.

1.7.1 Genome Sequencing

Sequencing refers to the use of certain techniques to determine the order of genetic material in either DNA or RNA or of amino acids in proteins. Automated techniques are now used which can quickly and accurately sequence complete genomes (Lerner and Lerner, 2008). Genetic sequences are the basis of bioinformatics research and many

databases and tools have been developed to organise and analyse these sequences to discover biologically relevant information.

1.7.1.1 Orthologs

One of the most exciting aspects of the plethora of complete genomes available is the ability to perform comparative analysis to find homologies, or a relationship of common descent, between genes. A homologous gene can be classified into two categories: an ortholog, genes related due to speciation or derived from a single common ancestor, and a paralog, genes related due to duplication, which can occur within one organism and/or between organisms (Koonin, 2005). The definition of ortholog has no direct meaning for functional characterisation of genes, but it does have implied meaning and oftentimes orthologs have equivalent function. One comparison between *E. coli* and *B. subtilis* for one-to-one orthologs (only one protein in each genome) did not produce one clear example of different functions between orthologs (Koonin, 2005). This does not hold true for organisms in different kingdoms, such as between bacteria and archaea or eukaryotes, which often show different functions. Additionally paralogs, though they may retain their ancestral function, can also display functional diversification and specialisation based on selective constraints. Lastly, although most orthologs have equivalent function, the reverse statement is usually false; many examples exist where non-orthologous and even non-homologous proteins perform equivalent functions. The functional equivalency of orthologs is often used to annotate genomes because it is impossible to experimentally derive the function of all genes in all sequenced genomes.

The characterisation of orthologs can be confounded in prokaryotes by the commonly occurring event called horizontal gene transfer (HGT), in which genes are acquired from an outside source; these genes could appear to be orthologous but are not true orthologs, and are rather called xenologs. Additionally, genes can appear to be orthologous when lineage-specific paralogous genes are lost, but these are in fact rather known as pseudoorthologs (Koonin, 2005).

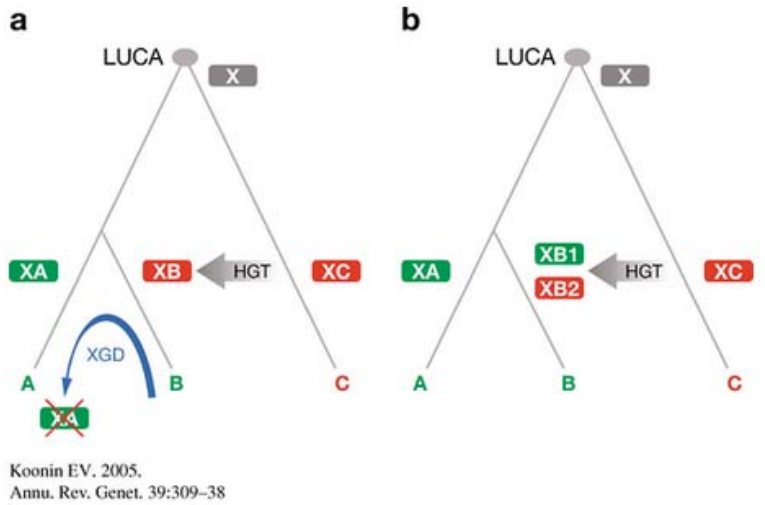


Figure 1.4 Pseudoorthology and xenoparalogy caused by horizontal gene transfer. Image (a) shows HGT leading to xenology, in which gene XB is transferred from species C and appears to be orthologous to XA, with XA being lost in species B due to displacement. Image (b) shows HGT leading to pseudoparalogy and pseudoorthology, in which XB2 has been transferred from species C and appears paralogous to XB1, and orthologous to gene XA (Koonin, 2005).

Xenologs and pseudoorthologs may appear to be orthologous but in fact are not, and thus have important effects on predicting gene function. Orthologs can be identified in multiple ways. First, orthologs are identified on the basis of phylogenetic analysis in which a gene tree is compared to the species tree and then are reconciled; HGT has an enormous effect on this method. Another method matches genes based on sequence and assumes that orthologous proteins are more similar to one another than to other genes in the genome and that matches are most likely formed between orthologs. This can lead to false negatives and false positives, particularly for genes that are paralogs, pseudoorthologs and xenologs. While these methods are by no means perfect, certain assumptions such as these are necessary due to the computing time needed to process whole genomes (Koonin, 2005).

1.7.2 Databases

With the development of all the sequenced genomes now available, the organisation of all this data has become essential. Thus a number of databases have developed to organise data and compile it into easily accessible locations for researchers to add to and utilise. There are a number of databases relevant to this study including the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (www.genome.jp/kegg/) (Aoki-Kinoshita and Kanehisa, 2007; Kanehisa et al., 2010; Ogata et al., 1999), Encyclopaedia of

Metabolic Pathways (MetaCyc) (www.metacyc.org) (Caspi et al., 2012), Integr8 (no longer operational) (http://www.ebi.ac.uk/integr8/), Universal Protein Resource (UniProt) (www.uniprot.org) (The UniProt Consortium, 2014) and Gene Ontology (GO) (www.geneontology.org) (The Gene Ontology Consortium, 2000).

1.7.2.1 KEGG

The early development of KEGG began in 1995 as part of the Japanese Human Genome Project (Ogata et al., 1999). The database stores and organises information in order to integrate genomic, chemical and systemic functional information (Kanehisa et al., 2008). It has been widely used as a reference for the biological interpretation of datasets such as those produced from sequencing (Kanehisa et al., 2006) KEGG is composed of 15 main databases including KEGG PATHWAY (pathway maps), KEGG ORTHOLOGY (KO) (manually defined ortholog groups), KEGG GENES (gene catalogues), KEGG ENZYME (information on enzyme nomenclature) and KEGG REACTION (metabolic reactions) (Kanehisa et al., 2010). Genome annotation is based on the KO system; using genes with known function, orthologous genes of organisms are assigned to orthology groups based on best-hit sequence matches. Once this annotation occurs, the genes with their new *K* numbers can be mapped to pathways (Kanehisa et al., 2012). KEGG contains 237 pathway maps, which are manually drawn graphical diagrams that show all the known reactions, no matter the organism or taxonomic group in which they function, related to a particular topic (for example phenylalanine metabolism). KEGG contains pathways that do not exist in MetaCyc such as xenobiotic degradation, glycan metabolism and metabolism of terpenoids and polyketides (Altman et al., 2013). KEGG also allows users to map their own data onto KEGG diagrams, for example colouring specific attributes (Kanehisa et al., 2012).

1.7.2.2 MetaCyc

MetaCyc is similar to KEGG in that it is a reference source that integrates metabolic pathways and can be used for pathway prediction of organisms based on their annotated genomes (Karp et al., 2002). It contains a searchable encyclopaedia of enzymes, enzymatic reactions, non-redundant metabolic pathways, substrate-level regulation, cofactor requirements, substrate specificity as well as a host of other information (Caspi et al., 2006). MetaCyc contains a large number of reactions

categorised into 296 superpathways, the MetaCyc equivalent of KEGG pathway diagrams (Altman et al., 2013). Additionally, all MetaCyc pathways are experimentally derived, providing a high level of accuracy and reliability (Caspi et al., 2014). The expected taxonomic range of each is also included in the database, allowing users to find pathways that are likely to exist in the organism being investigated. When a MetaCyc pathway is present for an organism then it is assumed that all the reactions within that pathway are present for that organism. Lastly, MetaCyc contains pathways that do not exist in KEGG such as those in actinobacteria, plants, fungi and metazoan (Altman et al., 2013).

1.7.2.3 Integr8

Integr8, though no longer in use, provided a single entry location to complete genome information (Kersey et al., 2005). It included all organisms with completely sequenced genomes at the time of database retrieval. Integr8 also allowed this data to be searched and for chosen components to be compared and extracted via downloads. Additionally, summary information for each genome could be accessed and downloaded for further analysis. The database included CluSTr, which hierarchically clustered proteins based on sequence similarity; this clustering proceeded via single-linkage clustering to create a similarity matrix, for which statistical significance was then assessed, creating a hierarchy of clusters (Petryszak et al., 2005). The CluSTr database formed the basis for the prediction of orthologs for each protein from the genomes. Within Integr8, the ortholog file for each sequenced genome could be downloaded; this file would include all predicted orthologs across all organisms in the database for that genome (Mulder et al., 2008).

1.7.2.4 UniProt

The goal of UniProt is to “provide a comprehensive, high-quality and freely accessible resource of protein sequences and functional annotation” (The UniProt Consortium, 2014). It compiles, interprets and standardises data from a variety of sources to provide a highly comprehensive catalogue of protein information. The UniProt Knowledgebase (UniProtKB) is entirely curated with both a fully reviewed and manually annotated section known as UniProtKB/SwissProt and non-reviewed and automatically annotated section called UniProtKB/TrEMBL. UniProt is the world leader in the provision of

comprehensive curation of literature-derive experimental data, and works with other databases to ensure that time and effort is not wasted on the duplication of work. Each page in UniProt is filled with the record of one gene and all its protein products in one organism and includes names, function, catalytic activity, pathways and all relevant information to those proteins (The UniProt Consortium, 2014). Thus, UniProt provide a highly accurate and reliable database of proteins along with corresponding relevant information regarding classification, function and activity.

1.7.2.5 Gene Ontology (GO)

The GO resource classifies products of genes with functional information by using structured and controlled vocabularies. This resource aims to provide a standard terminology with which to annotate genomes to enable consistent and accurate comparisons between information derived from a variety of sources. GO terms describe how (function) and where (process and location) gene products act and are composed of both the terms applied to a specific gene product as well as the evidence for the application of that term. Annotating gene products with GO terms proceeds by both manual (e.g. using literature sources) and automatic (e.g. using sequence similarity) methods. Terms are periodically reviewed for biological relevance in order to maximise utility and ensure complete coverage (The Gene Ontology Consortium, 2013).

1.7.3 Tools

One of the major tasks in bioinformatics is the development of user-friendly tools in order to help analyse and interpret large-scale sets of data such as whole genome sequences. These tools are often designed to make bioinformatics analyses possible for those without computational backgrounds such as biologists (Baxevanis and Ouellette, 2005). A number of tools relevant for this study include Basic Local Alignment Search Tool (BLAST) (blast.ncbi.nlm.nih.gov/Blast.cgi) (Camacho et al., 2009), Pathway Tools (http://bioinformatics.ai.sri.com/ptools/) (Karp et al., 2010) and MultiExperiment Viewer (MeV) (www.tm4.org) (Saeed et al., 2003).

1.7.3.1 BLAST

BLAST is a program used to perform sequence similarity analyses (Johnson et al., 2008). These sequence alignments are typically the first method used to connect newly sequenced DNA to sequences that have already been characterised. BLAST takes a

nucleotide or protein sequence and uses it to search across either a nucleotide or protein database to find matches (Boratyn et al., 2013). BLAST can be accessed either through the web interface or by download and operation on the command line, and uses a heuristic that first locates short matches between sequences before attempting alignments in order to shorten the required length of time to process the alignment(s). Once matches are found, BLAST provides statistics regarding the false-positive rate; the lower this rate, also known as the 'expect' value (e-value), the less likely the match is a false positive. BLASTP attempts to perform protein alignments between query sequences and target genomes. Query sequences can take the form of a sequence in FASTA format or sequence identifiers such as GenBank accession number (Ye et al., 2006). The e-value threshold for matches is often taken at 10^{-6}, with any values lower than this taken to signify a positive match; however, higher e-values might be applicable to find all possible positive matches or lower e-values might be necessary to sort orthologs and paralogs (Kerfeld and Scott, 2011).

1.7.3.2 Pathway Tools

Pathway Tools (PT) uses the MetaCyc database in order to predict metabolic networks based on genome sequence. PathoLogic uses 'key reactions' and taxonomic range (predicted in MetaCyc) to identify pathway memberships for a particular genome; 'key reactions' are those that are unique, or do not also function in other pathways, and are also defined in MetaCyc (Karp et al., 2011). The Cellular Overview tool in PT is used to show all the metabolic pathways for a Pathway/Genome Database (PGDB) in one integrated diagram similar to the Global Map in KEGG (Caspi et al., 2013). Additionally, PT creates a list of pathway holes, or situations in which a genome seems to lack the enzymes needed to catalyse reactions within a pathway; the database uses certain methods that combine evidence (such as homology, operons and context) to attempt to fill these holes, but remaining pathway holes are combined into a file available for download (Green and Karp, 2004).

1.7.3.3 MeV

MeV is actually a tool developed for the visualisation and analysis of microarray data (used to determine expression patterns in organisms). The tool allows users to organise microarray data into heat maps in which expression levels are represented by various

shades of two colours signifying the high and low values. It also allows user to analyse such data by utilising clustering and statistical tools (Saeed et al., 2006). These heat maps also allow for the visualisation and analysis of alternative types of data such as ortholog information across a wide variety of species.

1.8 Annotations

An annotation is composed of any type of additional information added on top of any existing data or document. Annotation incorporates additional layers of biologically significant knowledge to help classify and analyse large-scale data such as that generated by genome sequencing efforts (Gupta, 2009). Types of annotation for prokaryotes include structural, based on experimental evidence to find the physical characteristics of a gene, and functional annotation, based on sequence similarity to determine the function of the gene (Beckloff et al., 2012). Genome annotation is performed via two methods: the first is manual annotation, in which a researcher physically adds annotation information to a genome, and the second is automatic annotation, which takes advantage of various bioinformatics tools in order to automatically add information to the genome.

1.8.1 Automatic Annotation

Automatic annotation involves the automatic generation of annotations, usually based on sequence similarity using various bioinformatics tools. For raw genome sequences these automatic annotations typically output a set of open reading frames (ORFs) with start and stop codons, as well as predictions of function, metabolic pathways, gene ontologies and phylogenetic information based on sequence alignments (Tummler, 2010). A large proportion of these automatic annotations can be incorrect (14-58% in one study), meaning that results must be evaluated for accuracy and further work must be done to provide correct biological interpretations (Ederveen et al., 2013). Automatic functional annotation can be performed through the use of either pairwise or multiple sequence alignments, usually with phylogenetically related organisms. Functional annotations are assigned to genes based on alignments with already-annotated reference genomes (Pirovano and Heringa, 2008). Automatic annotation is more efficient and less labour intensive, but usually requires manual annotation to improve accuracy and maximise biological analyses.

1.8.2 Manual Annotation

Manual annotation involves the review and curation of automatically generated annotations by human experts (Richardson and Watson, 2013). This typically makes great use of the literature and assigns more specific information to genes based on direct experiments on those genes and gene products and/or their orthologies with genes characterised in substantial depth in the literature. When little functional data is located, a more in depth analysis in performed via sequence analyses, database mining and literature searches. These annotations will either create or improve annotations or allow a hypothesis about the function to be generated (Tummler, 2010). Manual annotation will improve reliability and biological significance of annotation but is also time consuming. Most genome annotation processes involve a combination of both automatic and manual annotation to maximise accuracy and efficiency (Richardson and Watson, 2013).

1.9 Metabolic Pathways

A metabolic pathway is a "series of enzyme-catalysed chemical reactions occurring within an organism, in which a principle chemical is modified" (Caspi et al., 2013). They are the true functional units of metabolic systems and enable an organism-wide interpretation of cellular activities (Schilling et al., 2000). Pathway data includes three levels of information including the metabolites that form the basis, the reactions, which are built on metabolites, and pathways, which are composed of reactions (Altman et al., 2013). Metabolic pathway networks can provide information on organism response to different conditions and, for pathogenic bacteria, help elucidate disease progression and interdependent mechanisms between host and microbe; the more complete a metabolic pathway is, the greater the biological relevance of its analysis and interpretation (Papin et al., 2003). To enable analyses based on metabolic pathways, they first must be classified, and then they must be compared to identify biologically relevant discoveries.

1.9.1 Functional Classification

Functional annotation and classification of genes is usually accomplished via sequence alignments with highly characterised reference genomes such as *E. coli.* Accuracy of pathway prediction is heavily based on the coverage of the reference pathway of the

database. MetaCyc and KEGG share an estimated 3,600 reactions, leaving about 1,000 reactions unique to each database. Pathway predictions performed using KEGG have been theorised to be less accurate than for MetaCyc, because MetaCyc provides additional reaction attributes such as taxonomic range (Altman et al., 2013). Functional annotation based on homology is a subject of debate, but multiple studies have shown the functional equivalency of orthologs; while the fundamental functions of orthologs sometimes change, these changes are rare and often associated with major evolutionary transitions involving a significant acceleration of evolution (Koonin, 2005). It is usually assumed that paralogs are less likely to retain function and therefore have lower levels of functional equivalency (Altenhoff et al., 2012; Chen and Zhang, 2012; Thomas et al., 2012); however, there is also some evidence to the contrary in that paralogs might have just as much if not more functional equivalency as orthologs (Nehrt et al., 2011). In all, the ortholog conjecture seems to be supported, and it is likely that paralogs have lower levels of functional equivalency as compared to orthologs. Since basic sequence alignments do not distinguish between orthologs and paralog but simply analyse homology between sequences, the potential effects of paralogs must be taken into account.

1.9.2 Comparison of Pathways

The development of metabolic pathway networks for sequenced genomes has enabled the comparison of these networks. For example, comparisons of genome-scale metabolic pathway reaction content and networks have been used to perform phylogenetic analyses on organisms (Hong et al., 2004). Additionally, these comparisons can help identify virulence factors that explain why certain organisms or strains of organisms are more virulent during infection of hosts (O'Callaghan and Stebbins, 2010). Network comparisons have also been performed in order to locate differences in metabolic phenotype. These network comparisons can also be used for many other purposes such as multi-species studies and host-pathogen interactions. As of 2009 these last two uses of metabolic reconstructions were relatively represented the least in the literature but can provide biologically relevant predictions of phenotype and bacterial activity (Oberhardt et al., 2009). Lastly, many studies make use of mycobacterial model systems such as *M. smegmatis, M. bovis* BCG and *M. marinum;* comparative analysis of genome-scale metabolic networks can assist in identifying

pathways that are both similar and divergent between the model organisms and Mtb (Shiloh and Champion, 2010).

1.10 Pathway Holes

A pathway hole, or 'missing' pathway, is a metabolic reaction within a pathway for which no catalysing enzyme has been located in the genome. Many possible pathway holes result from situations in which genes coding these enzymes do exist but have simply not been identified by the annotation methods, especially when one gene has multiple functions (Karp et al., 2010). Alternatively pathway holes can signify occasions when the bacteria might utilise, or hijack, the host metabolism in order to acquire certain nutrients and accomplish certain metabolic needs. Bacteria belonging to *Chlamydiae*, obligate intracellular pathogens that replicate within vacuoles, have been shown to redirect vesicles and hijack organelles in order to ensure the acquisition of essential nutrients (Saka and Valdivia, 2010). The identification of genuine pathway holes can therefore provide important information regarding potential host-pathogen interactions.

1.11 Need for New Drugs

New drugs are urgently needed to address some of the concerns outlined above (Beste and McFadden, 2013). These drugs have several requirements in order to help end this epidemic of tuberculosis. First, new drugs are required that can combat Mtb during both active and latent infection. Next, new drugs must be effective against resistant strains of Mtb, meaning that they must have novel targets compared to currently used drugs. Third, they must be able to treat patients with HIV coinfection, and thus have no interactions with antiretroviral drugs. New drugs must also be able to shorten the duration of treatment to reduce issues with patient adherence to treatment regimens. Additionally, these drugs should be low-cost options in order to allow patients in resource-poor regions to have access to new treatment options. Lastly, new drugs do not necessarily need to function on their own as monotherapy but instead should be included in new, comprehensive regimens that accomplish all of the functions thus described (Ma et al., 2010).

1.12 Thesis Rationale

Knowledge of the metabolism of Mtb will increase understanding of the disease as it develops within hosts and greatly assist in the rational development of drugs. With a greater understanding of the metabolome, or complete metabolism, doctors and researchers will better understand the course of the disease, how to quickly diagnose TB and better treatment regimens. One of the major steps in the drug development pipeline involves the identification and validation of appropriate targets, which is presently a major bottleneck. Many currently used drugs have been developed without the knowledge of targets, possibly because of the inability to use standard methodologies to identify these targets on a large scale (Raman et al., 2008). Additionally, many drugs have bad side effects caused by polypharmacology, or when drug molecules interact with multiple targets (Reddy and Zhang, 2013). Comparing pathway and reaction information over a broad spectrum of organisms (phylogenetic profile) can assist in finding these new drug targets without affecting proteins in the host or beneficial bacteria residing in the host. Proteins of Mtb without orthologs in other organisms can identify enzymatic reactions, which, if attacked, would not affect the metabolism of host cells nor beneficial bacteria within the host. By finding these proteins, potential drug targets can be identified for future drug development. Additionally, by using a phylogenetic profile to identify Mtb proteins with and without orthologs, essential pathways can be detected by observing orthologs in *Mycobacterium leprae*, a species with a highly reduced genome. Furthermore, by comparison with an anaerobic bacterium, potential pathways essential for anaerobic metabolism can be identified, thereby finding potential drug targets for the persistent phase of infection. Lastly, mapping these pathways can help identify 'missing' pathways or pathway holes, possible circumstances in which Mtb uses the host metabolism to accomplish its metabolic needs.

1.13 Aims of the Study

This study consists of three main steps:

i. Completing the metabolic map by adding novel functional annotations to the genome of *M. tuberculosis* H37Rv
ii. Identifying reactions and pathways of interest using the phylogenetic profile
iii. Identifying pathway holes, or missing pathways

1.14 Road Map

The remainder of the thesis is composed of the following chapters:

- **Chapter 2** details the materials and methods used to complete the aims of this study. This includes the steps taken to add manual annotations to the genome based on both functional terms such as GO ontology and using KEGG mapping diagrams as a basis for protein matching with BLASTP.
- **Chapter 3** details the results of the methods described above; this includes the additional annotations and creation of tables recording pathway holes.
- **Chapter 4** discusses the results. First the possible confounding factors of the results are considered. Secondly the results of the annotations are examined, with further examination and mapping of three KEGG pathway diagrams with many newly characterised individual reactions and additional branches of pathways. Next the phylogenetic profile is analysed to discover important pathways and reactions, which include those in which Mtb shares both few and many orthologs with *M. leprae,* few orthologs with *E. coli* and *H. sapiens* and many orthologs with *C. glutamicum*, a facultative anaerobe. Lastly, pathway holes are considered to find those for which Mtb may use the host metabolism to accomplish its metabolic needs.
- **Chapter 5** reviews the important results found and summarises the conclusions drawn from these results. It evaluates the lessons learned and future research that must be done to advance the provided results.

2 Materials and Methods

2.1 Extraction of Ortholog Data

The Mtb ortholog data was downloaded from Integr8, which has since been replaced by Ensembl genomes (Kersey et al., 2005). The ortholog .txt files were downloaded for each of five different strains of Mtb on 21 May 2012. These five strains include ATCC 25618 (H37Rv), ATCC 25177 (H37Ra), Oshkosh (CDC1551), F11 and KZN 1435. Each of the ortholog files show a list of all proteins encoded by the genome in the strain, along with all the known orthologs of that protein in all other species existing in the database at that time. Each ortholog entry includes the ortholog accession number, the species name and the taxonomic information.

Information from these five ortholog files was then extracted and reorganised into a phylogenetic profile, along with pathway data downloaded from KEGG (Kanehisa et al., 2010). To do this, a script was written using Python (http://www.python.org/) to create a matrix with the GenBank accession number, EC number and pathway in the left-hand columns and the ortholog information filling in the rest of the columns. In this matrix, a '1' indicates that an ortholog is present and a '0' denotes that an ortholog is absent. An example of this structure is shown in Table 2.1.

Table 2.1 Example structure of the compiled matrix. This shows the accession number of each protein, along with the strain it has been derived from, its pathway membership, EC number and any orthologs it has across the phylogenetic profile.

Accession Number	Strain	Pathway Level 1	Pathway Level 2	Pathway Level 3	EC Number	Org 1	Org 2….
P12345	H37Rv	Metabolism	Energy Metabolism	Methane Metabolism	EC:1.2.3.4	1	0
P67890	KZN	Metabolism	Carbohydrate Metabolism	Glycolysis/ Gluconeogenesis	EC:5.6.7.8	1	1

To create this matrix, the Python script first processed the H37Rv ortholog file, extracting all the proteins and their orthologs. After running through H37Rv, the script ran through the rest of the strains in the order of H37Ra, F11, CDC1551 and lastly KZN, extracting only those proteins that were not orthologs of H37Rv. These 'unique' proteins and their orthologs were then appended to the matrix. Thus, the final matrix includes all H37Rv proteins and those proteins of each strain that were not present in the previously extracted strains. Theoretically, all of these proteins should be 'unique'

i.e. not orthologous to one another. This was done to retrieve a complete list of all possible Mtb proteins.

Pathway information was previously derived from a dataset created by a colleague in the group using the KEGG database. This file contained information for *M. tuberculosis* H37Rv regarding the protein accession number, EC number(s), pathway(s) and functional description. Pathway membership and EC numbers were extracted from this dataset and added to the matrix according to their protein accession number. Each pathway is described by terms at three different levels, becoming more specific by the third level. Thus columns were created for each pathway level in the final matrix, as shown by the example above.

Many of the proteins with KEGG pathway information included multiple pathways per protein. Thus, when creating the matrix, each protein (and corresponding ortholog data) with multiple pathway membership was duplicated so that those proteins filled as many rows as they belonged to pathways. If the protein had more than one EC number then all were included for each protein entry.

2.2 Incorporating Additional Orthologs

Once creation of the structure of the phylogenetic profile was completed, it was noticed that some proteins showed no orthologs when in fact there was a known ortholog in that strain or species. It was concluded that the Integr8 data must have used strict conditions when identifying orthologs in other species and strains, thus missing some orthologs. To remedy this, a reciprocal BLASTP was used to find additional orthologs in Mtb strains H37Rv, H37Ra, F11 and KZN using CDC1551 as the reference (Boratyn et al., 2013).

The additional ortholog data was then incorporated into the matrix. Anytime a '0' showed in the matrix for any of these strains, but the additional dataset showed a '1', that entry was changed to '1' in the matrix. After completing this, the matrix now signified that an ortholog is present when there was an ortholog shown in either the original Integr8 data or the data obtained through the reciprocal BLASTP. This applies only to those five strains of Mtb: H37Rv, H37Ra, CDC1551, F11 and KZN.

2.3 Removal of Orthologous Proteins in H37Ra, F11, KZN and CDC1551

The same strict conditions of Integr8 also mean that some of the added 'unique' proteins for H37Ra, CDC1551, F11 and KZN, were in fact not unique. Thus, these proteins were removed from the matrix. To do this, first a search was performed on the UniProt website using the GenBank accession number. All UniProt entries of the F11 proteins included a link to the corresponding EMBL ENA (European Nucleotide Archive) webpage (www.ebi.ac.uk/ena/) for the gene encoding that protein (Leinonen et al., 2010). For some of these, the ENA webpage included information about which H37Rv protein that F11 protein was mapped to. In these cases, the matrix was checked to make sure that the H37Rv protein was already included. If it was already present, then the F11 protein was removed from the matrix. EMBL ENA webpages for the proteins of the other three strains of Mtb (H37Ra, KZN and CDC1551) did not show mapping information. Therefore, once a UniProt search was performed for each of these proteins, the ordered locus name was used to perform a search on the KEGG website. In some cases, KEGG provided no additional information about that protein, often saying only that it was a hypothetical protein. When this occurred, the protein was left as is in the matrix. In other cases, KEGG showed an orthology identifier. In these instances, the KEGG orthology number was used to find which H37Rv protein, if any, was its ortholog. When the protein showed an ortholog in H37Rv, and that H37Rv protein already existed in the matrix, then that protein from the other strain was removed from the matrix. Therefore, this left only unique proteins for each of the five strains in the phylogenetic profile.

2.4 Manual Annotation of Proteins without Pathway Data

To increase pathway data coverage, the remaining proteins were manually annotated using the UniProt (www.uniprot.org) and KEGG websites (The UniProt Consortium, 2014). Even though the KEGG pathway data was originally obtained from a colleague, it seems this data was incomplete, and many new annotations could be added from the KEGG database. In addition, the proteins added from the H37Ra, F11, KZN and CDC1151 strains had as of yet no pathway information and were thus also manually annotated. A column was added to the matrix in order to record additional information such as gene name, function, Gene Ontology (GO) terms and other miscellaneous information. The

following account describes the steps taken for each protein to manually add pathway, EC number and functional data.

2.4.1 UniProt Search

The first step towards the addition of pathway data and EC numbers involved performing a search in UniProt. The UniProt entries provided information on the gene name, gene function and GO terms. This information was copied into the 'Additional Information' column. In some cases, the 'General Annotation' section included a note that this protein was a high-confidence drug target. In these cases, the row of that protein was highlighted yellow. When a protein had an EC number, then that number was added to the matrix and the box was coloured blue to indicate the source. The UniProt page also showed the ordered locus name, which could then be used for further enquiries.

2.4.2 KEGG Search

Using the ordered locus name found in UniProt, a KEGG search was performed for each protein without pathway data. Some proteins displayed both EC numbers and pathway membership information. In these cases, all pathways were added to the matrix, again duplicating the row of the protein as many times as the number of pathways to which it belongs. For proteins that already had EC numbers found on UniProt and for which the KEGG EC number was the same, nothing was changed. For proteins that did not have EC numbers on UniProt, the KEGG EC number was added to the matrix and the cell was coloured green to signify the source. For proteins that showed different EC numbers on UniProt and KEGG, both numbers were added to the matrix, with the UniProt EC number displayed first. For some proteins KEGG displays only pathway information or EC numbers. In these cases, the available information was added to the matrix.

2.4.3 KEGG BRITE Hierarchies

KEGG BRITE is a collection of functional hierarchies using structured vocabularies and can be used to represent functional information (Kanehisa et al., 2008). For some proteins, KEGG would provide no pathway information but would have assigned BRITE terms to a protein. These BRITE terms were included in the reference hierarchy found on KEGG Orthology (KO) (http://www.genome.jp/kegg-bin/get_htext?ko00001.keg), and thus showed three levels in the same manner as the pathway terms. Many proteins

displayed BRITE terms only with no pathway information, and these BRITE terms would be included in the pathway cells for that protein. Examples of these BRITE terms include 'Glycosyltransferases', 'Translation Factors', 'Transcription Machinery', and 'Peptidases', among others. This was performed to enable functional classification of as many genes as possible, and so that the number of remaining unannotated genes needing further investigation would be reduced.

2.4.4 Converting UniPathway Pathways Into KEGG Pathways

UniProt provided pathway information for some proteins that did not have information on KEGG. In these cases, the pathways were linked to UniPathway (http://www.grenoble.prabi.fr/obiwarehouse/unipathway), which does not use the same naming structure as KEGG. On the UniPathway website some pathways show cross-references with KEGG or with MetaCyc pathways. If the pathways were cross-referenced with KEGG pathways, then those KEGG pathways were added to the matrix. Though some did not show KEGG cross-references, their KEGG pathway could be inferred due to name similarity or specific terms and would also be added. For a few pathways, no appropriate match could be found in KEGG and thus these pathways were added to the matrix in the UniPathway naming structure. For example, protein O69670 showed no pathway membership in KEGG but was assigned to the 'amino-acid biosynthesis; ergothioneine biosynthesis' pathway in UniProt. When this pathway was found in UniPathway, there was no KEGG mapping information available (Figure 2.1). Therefore, the UniPathway terms were added to the matrix without conversion.

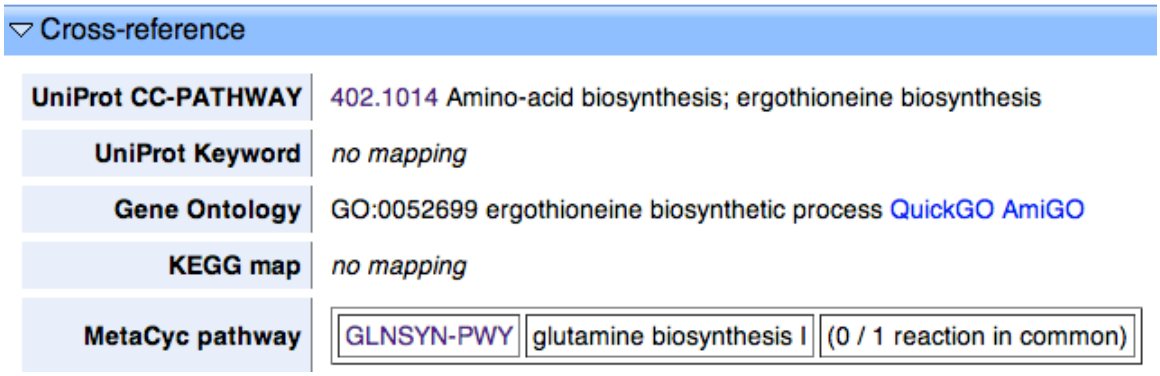


Figure 2.1 Example of UniPathway pathway membership information. This image shows the pathway membership for a protein shown in UniPathway as acting within ergothioneine biosynthesis but for which no KEGG pathway membership is available (Morgat et al., 2012).

Although no KEGG mapping data is provided, KEGG terms could sometimes be derived via another method. On the 'Overview' tab of these pathways there is a 'Pathway hierarchy: IsA relationships' dropdown menu. By clicking on this dropdown menu, a tree view of the pathway hierarchy is shown, with both UniPathway and GO terms. In this tree view, some of these pathways were nested in pathways that directly corresponded to KEGG pathway terms. For example, the previous pathway 'ergothioneine biosynthesis' is shown in the tree view (Figure 2.2).

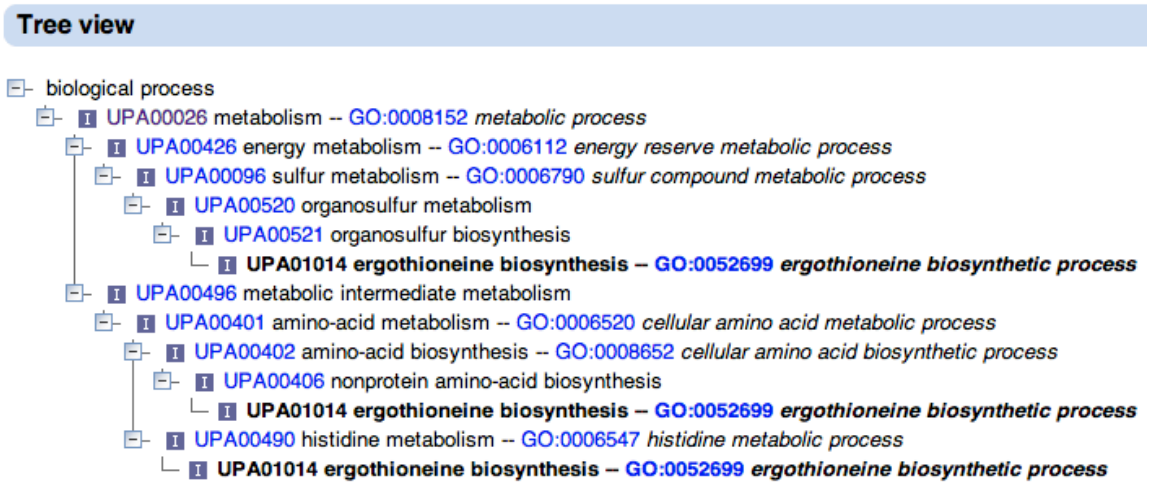


Figure 2.2 Pathway membership tree view in UniPathway. These nested pathway trees could be used to locate equivalent KEGG pathways and assign annotations to certain genes (Morgat et al., 2012).

In this case 'ergothioneine biosynthesis' is nested within 'sulfur metabolism' and 'histidine metabolism', which are both KEGG terms. Thus these two pathways were also added in the matrix for proteins in the 'ergothioneine biosynthesis' UniPathway pathway.

2.4.5 Using EC Numbers to Find Additional Pathway Memberships

Many proteins were assigned EC numbers on UniProt and/or KEGG but did not show any pathway memberships. In these cases a search was performed on KEGG Enzyme for that EC number. From the EC number's webpage is a link to all KEGG reactions that are associated with that EC number. Once on the webpage for all associated reactions, it is possible to scroll through to see in which pathway(s) each reaction is involved. All possible pathways were added to the matrix for that protein, even though some undoubtedly did not apply to Mtb. The aim was to include all possible pathways in order to not miss any potential memberships. Only certain pathways, such as 'Photosynthesis' or 'Insect Hormone Biosynthesis' were excluded. Later on these

memberships would be investigated further in order to determine the veracity of memberships derived through this method.

2.4.6 Using GO Terms and Gene Names to Derive Additional Pathways

UniProt provides a list of applicable Gene Ontology (GO) terms for each protein. In some cases, the GO terms for molecular function or biological process directly signified certain pathway membership. For example, a GO term might be 'histidine metabolism' and that pathway would be added to the protein's pathway memberships.

In other cases the GO term would signify a function that did not directly correlate with KEGG pathways. In this instance the reference hierarchy found on KEGG Orthology (KO) would be used to find proteins with the same function. By searching for keywords from the function of the protein, the protein name or from the GO terms, similar proteins could sometimes be found in the reference hierarchy. When matching proteins or orthologies were found, whichever pathways it belonged to would be added to the matrix.

2.4.7 Additional Pathways from UniPathway

In order to check if any of the still uncharacterised proteins had UniPathway data that had been missed, or could not be converted into KEGG pathway mapping, the complete list of all 424 proteins with pathway data on UniPathway was downloaded. This list was then cross-referenced with all proteins without pathway data in the phylogenetic profile. All proteins with UniPathway pathway membership but without KEGG pathway membership were highlighted. Then these proteins were manually searched on the UniPathway website. These pathways had no equivalent pathway in the KEGG pathway mapping system, and so the pathways were recorded as they are found on the UniPathway website. Additional pathways added included 'cell wall polysaccharide biosynthesis,' 'coenzyme F0 biosynthesis,' 'molybdopterin biosynthesis,' 'trehalose degradation,' 'lipoprotein biosynthesis' and 'mycolic acid biosynthesis.'

By adding UniPathway membership data only to those proteins without KEGG pathway membership data, the pathways were not complete as shown on UniPathway. Thus, all proteins that belong to each added UniPathway pathway were found and then added to the matrix, duplicating proteins that already had KEGG pathway data.

2.5 Reordering of Matrix

With all these proteins, EC numbers and pathway information added to the matrix, the next step was to reorder the matrix. All proteins were reordered to cluster them into pathways with proteins having EC numbers but no pathway data beneath, followed by proteins without any EC number or pathway information at the bottom.

2.6 Mapping Proteins to KEGG Pathway Diagrams

All proteins on the list characterised by pathways were then mapped to the KEGG pathway diagrams. Each protein already existing in the KEGG database was marked on the diagrams. Then any additional proteins that were identified as enzymes of particular reactions were also mapped onto the pathways. This included proteins that were duplicates of reactions as well as proteins not already mapped into pathways.

2.7 Deriving Additional Pathways Using BLASTP

The next step was to further characterise proteins without pathway information into unidentified KEGG reactions. Firstly, KEGG maps of each metabolic pathway were compared between *M. smegmatis,* Mtb and *M. leprae*. The differences, in terms of reactions filled in by proteins, between the pathways were noted. Since *M. smegmatis* usually had more characterised reactions, these were then used to find additional information about the metabolome of Mtb. For each of the *M. smegmatis* proteins identified as belonging to a KEGG reaction, the NCBI-GI (genInfo identifier) accession number was imported into BLASTP in order to find matches with Mtb proteins. A cut-off e-value of 10^{-6} was used in most cases; in some cases where no results could be found, the cut-off value was increased to 10^{-3}. All sequences with a BLASTP hit were recorded in a table along with sequences for which no results were found (can be found on the CBIO website). These matched proteins of Mtb were then mapped onto their reactions for each of the metabolic pathway maps. The BLASTP results include up to 100 matches (for all Mtb strains) per sequence up to the cut-off value; all hits within these limits were recorded. All results were then added to the pathways both in the matrix and on the pathway maps.

2.7.1 Using BLAST on Other Organisms

Each pathway map for H37Rv was compared with other closely related organisms in order to annotate additional proteins with BLASTP matches and thus fill in pathway

holes. For each metabolic pathway, maps for *M. ulcerans, M. vanbaalenii, M. marinum M, Corynebacterium glutamicum ATCC 13032 (Kyowa Hakko), Nocardia farcinica, Rhodococcus jostii RHA1, Rhodococcus erythropolis PR4, Gordonia bronchialis* and *Streptomyces coelicolor* were compared in the order thus stated. For reactions that were not yet mapped to proteins in Mtb but showed proteins in any of the other organisms, NCBI-GIs of these proteins were searched against proteins in Mtb and mapped to the pathways when results were found. These BLASTP matches also used the cut-off e-value of 10^{-6}. Identified proteins were then added to their respective pathways in the matrix. The annotated proteins of three pathways were then evaluated using Ensembl Genomes (www.bacteria.ensembl.org) to identify possible operons between them and previously annotated proteins (Flicek et al., 2014). The genes encoding these proteins were mapped to their location on the Mtb chromosome in order to see if they are situated next to each other.

2.8 Analysing the Phylogenetic Profile

MeV (http://www.tm4.org/index.html) was used to visualise the phylogenetic profile across all organisms as a heat map (Saeed et al., 2003). The resulting image was quite large and so the total number of proteins for each organism in each pathway were used rather then every single protein in *M. tuberculosis* H37Rv. After visualising (as a heat map in MeV) the phylogenetic profile, it was observed that a number of organisms showed very incomplete data; they displayed either none or extremely few orthologs with Mtb. Thus, these species were removed from the phylogenetic profile. Removal was performed based on the total number of pathways. For example, any organism that did not have at least one ortholog in at least 50 of the 99 metabolic pathways contained in the profile was removed. This way only those organisms that were relatively well characterised were included. To further reduce the size of the heat map all species with multiple strains or variations were removed so that only one organism for each species remained in the heat map. Once, this was completed the proportion of orthologs for each pathway in each organism was calculated by dividing the number of orthologs in that pathway for that organism by the total number of *M. tuberculosis* H37Rv proteins in that pathway. Then the average proportions of all the organisms for each pathway was calculated in order to rank pathways by most to least conserved across the phylogenetic profile.

2.9 Deriving the Pathway Summary Totals for *Escherichia coli*

In order to compare the pathway summary of Mtb, the pathway summary of *E. coli*, one of the most well annotated genomes available, was derived from KEGG Orthology. For this the *E. coli* strain K-12 MG1655 was chosen. All proteins within each pathway were counted and the totals were entered into the matrix next to the totals from Mtb and *M. leprae* for comparison.

2.10 Identifying 'Missing' Pathways or Pathway Holes

Multiple methods were used to identify possible 'missing' pathways or pathway holes within the maps. A table of all possible 'missing' pathways was created which shows the missing reaction along with the evidence identifying it as a pathway hole (Appendix C).

2.10.1 Pathway Tools

First, the pathway holes file was downloaded from version 16.5 of Pathway Tools (Karp et al., 2011). This file includes pathway holes for *M. tuberculosis* H37Rv as predicted by Pathway Tools. For each possible pathway hole, the reaction and EC number is identified. These reactions were then located on the pathway maps and marked as possible holes needing to be filled.

2.10.2 KEGG

The global map for *M. tuberculosis* H37Rv was observed to obtain additional evidence of pathway holes. On the global map, some reactions were shown in colours while others were grey. Those in grey seemed to be reactions that were believed not to exist in Mtb. Of the coloured reactions, some were shown in dark colours and others in pale colours, signalling that the reactions probably did exist but had not been mapped to particular proteins within the genome. This evidence was not taken as conclusive but was used in conjunction with other methods to identify pathway holes.

2.10.3 Existence in Closely Related Organisms

If reactions were observed to exist in closely related organisms but no matches were found using BLASTP, then this could possibly serve as signifiers of pathway holes. This evidence was also not taken as conclusive but used in conjunction with other possible indicators of pathway holes.

2.10.4 Resemblance

The final indicator for the identification of a reaction as missing or as a pathway hole was the existence of characterised reactions on either side of the reaction. For example if a reaction could not be mapped to a certain protein in Mtb but the preceding and succeeding reactions have been mapped to proteins, then this is taken as evidence of a hole. This includes reactions where the immediately adjacent reactions were mapped as well as reactions in which the adjacent reactions two and three steps away were mapped. Reactions of this type were observed for each of the pathways and then included on the chart of missing pathways. An example of one of these reactions is shown in Figure 2.3.

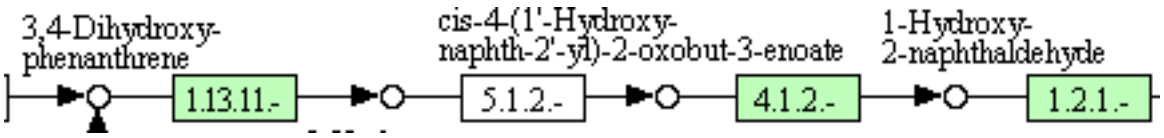


Figure 2.3 Example of a pathway classified as 'missing' or a pathway hole. The evidence for this classification is due to the characterised reactions on either side of this reaction, with only the one step missing in between. This reaction was thus added to the missing pathways file.

2.11 Conclusion

In this way we integrated data from a variety of sources to attempt to complete the metabolome of Mtb, prepare the data for comparison with other organisms and identify pathway gaps, such as pathway holes.

3 Results

3.1 Extraction of Ortholog Data

3,887 proteins were extracted from the original H37Rv ortholog file, with 294 additional proteins from H37Ra, 574 from F11, 56 from CDC1551 and 384 from KZN. In total, the original extracted phylogenetic profile thus comprised 5,195 proteins from all five strains of Mtb. All of these proteins should have been unique, or not orthologous to one another, but some were in fact orthologs and will be discussed later.

A total of 1,161 organisms were included in the final phylogenetic profile. This includes all organisms in the downloaded Integr8 files that show at least one ortholog with any of the five strains of Mtb. Some organisms included are *Homo sapiens* (humans) with 595 orthologs shown, *Escherichia coli* MG1655 with 789 orthologs shown and *Streptomyces coelicolor* with 1,014 orthologs. The five strains of Mtb included display orthologs for only 66% of the total 5,195 proteins on average. Therefore, many orthologous proteins were clearly missed using this dataset.

The pathway data obtained from a colleague included pathway membership and EC number information for 855 *M. tuberculosis* H37Rv genes. Upon creation of the initial matrix, pathways of these proteins were incorporated, leaving a substantial number of genes without pathway information or EC numbers.

3.2 Incorporating Additional Orthologs

As mentioned before, the five strains of Mtb show orthologs for only about 66% of genes (using Mtb H37Rv as reference). After incorporating the additional data from reciprocal BLAST results to find additional orthologs, many new orthologs were added, as shown in Table 3.1.

Table 3.1 This shows the number of orthologs for the five strains of *M. tuberculosis* in the original and updated phylogenetic profile.

Number of Orthologs for Each Strain	KZN 1435	F11	ATCC 25618	ATCC 25177	CDC1551
Number of Orthologs in Integr8 Dataset	3,340	3,189	3,475	3,523	3,626
Number of Orthologs With Added BLASTP Dataset	4,008	3,827	4,022	3,889	3,812

Therefore, a substantial number of orthologs were added for each of the five strains of Mtb, resulting in a more complete profile of orthologous proteins.

3.3 Removal of Orthologous Proteins in H37Ra, F11, KZN and CDC1551

This step is undertaken to ensure that all orthologous proteins should be removed from the dataset, leaving only unique proteins for the five strains of Mtb. Removing proteins that actually do have orthologs by using UniProt and KEGG orthology groups reduced the number of proteins in the phylogenetic profile. A comparison of the numbers of unique proteins before and after removal is shown in Table 3.2.

Table 3.2 Table showing the number of unique proteins from each strain in the phylogenetic profile. It compares the number of unique proteins as extracted from the original Integr8 dataset with those remaining after the data has been cleaned.

Number of Unique Proteins for Each Strain	H37Rv	H37Ra	F11	CDC1551	KZN	Total
In Original Dataset from Integr8	3,887	294	574	56	384	5,195
After Manual Removal	3,887	109	45	39	99	4,179

Many duplicate proteins were therefore removed from the phylogenetic profile, and the remaining dataset should include only 'unique' proteins. On the other hand, some proteins that exist in the other strains might not be included. For example, the genome of CDC1551 has 4,202 proteins in its complete proteome in UniProtKB while only 3,812 of the proteins for this strain were in the ortholog file. Since we were primarily focusing on the well-characterised laboratory strain, H37Rv, we were not concerned about this.

3.4 Manual Annotation of Proteins without Pathway Data

A number of proteins were added to the downloaded KEGG pathways by manual annotation using information derived from UniProt, KEGG and UniPathway. These include proteins that can fill current 'pathway holes' as well as proteins that can be

assigned to reactions that have multiple enzymes within the genome that catalyse that same reaction. By using these various sources and databases, an additional 553 proteins were annotated and characterised into pathways.

3.4.1 UniProt Search

While the UniProt search itself did not add any pathways, it did allow for the addition of many EC numbers. In total, EC numbers were added to 1,115 proteins based on UniProt searches. Some of these EC numbers were not found in KEGG and were thus added to proteins in the phylogenetic profile. In addition, proteins could be added to some pathways based solely on EC number.

3.4.2 KEGG Search

KEGG searches were then performed on remaining proteins without pathway data. A total of 354 EC numbers (which were not included in the original dataset) were added using this method. For some reason some of these proteins belonged to pathways but were not shown in the original pathway dataset obtained; therefore, these pathways were added to the matrix.

3.4.3 KEGG BRITE Hierarchies

While KEGG showed no pathway information for many proteins, some of these proteins did display BRITE hierarchies. Added BRITE hierarchies include enzyme families, glycosyltransferases and polyketide biosynthesis proteins, among others. For proteins without pathway information but with BRITE hierarchies, these hierarchies were added to the matrix. A total of 78 proteins were classified on the basis of KEGG BRITE hierarchies.

3.4.4 Converting UniPathway Pathways Into KEGG Mapping Pathways

Some proteins could be added to pathways based on their UniPathway mapping hierarchy. These pathways were added to the KEGG pathways by matching terms between the two databases. These include those belonging to the citrate cycle (TCA cycle), glycerophospholipid metabolism and porphyrin and chlorophyll metabolism, among others. In sum, 14 proteins were added into pathways based on their UniPathway pathway memberships.

3.4.5 Using EC Numbers to Find Additional Pathway Memberships

As mentioned previously, many EC numbers could be added to proteins in the matrix based on the UniProt and KEGG searches. Some of these EC numbers were found in pathways, and so some proteins could be added to pathways based solely on EC number.

3.4.6 Using GO Terms and Gene Names to Derive Additional Pathways

Using GO terms to map proteins to pathways allowed the characterisation of 49 additional proteins into many different metabolic pathways. The GO terms could be any level of specificity as long as they matched groups within KEGG ORTHOLOGY. These pathway assignments were then confirmed or refuted, depending on whether they could be mapped to specific locations in the pathways. GO terms were also often used in conjunction with EC numbers in order to determine functional characterisation. In many cases, protein pathway assignments based on GO terms were later refuted, but 49 proteins were confirmed and mapped to specific pathways.

3.4.7 Additional Pathways from UniPathway

Some UniPathway pathways had no equivalent pathway in KEGG, and so these were added to the matrix on their own. These include ergothioneine biosynthesis, L-arginine biosynthesis, L-cysteine biosynthesis, cell wall polysaccharide biosynthesis, coenzyme F0 biosynthesis, molybdopterin biosynthesis, quercetin degradation, trehalose degradation, mycolic acid biosynthesis, polypeptide chain elongation and lipoprotein biosynthesis. In all, 67 proteins were annotated with pathway information derived from the UniPathway database. Although these could not be mapped to KEGG pathway diagrams, they reduced the number of uncharacterised proteins requiring further investigation. Table 3.3 shows the numbers of proteins annotated by the manual annotation methods described thus far. Since many proteins function in multiple pathways, these numbers do not reflect the number of unique proteins added.

Table 3.3 Table showing the numbers of proteins annotated by each manual annotation method. The H37Rv proteins that were annotated are often duplicates, meaning they have been assigned to multiple pathways and so the counts are higher.

Annotation Method	Number of Annotated Proteins (includes proteins belonging to multiple pathways)
Original KEGG Dataset	2,766

UniProt EC Numbers	33
EC Number from KEGG	17
KEGG BRITE	78
Converted UniPathways	14
GO Terms	49
UniPathway Pathways	67
Total Functionally Annotated Proteins before BLASTP	3,024

Lastly, 323 EC numbers were added to proteins that could not be mapped to pathways in KEGG. Of these 39 EC numbers came from the original KEGG pathway data used to create the profile, 46 EC numbers came from the manual annotation using KEGG and the EC numbers for 238 proteins were derived from UniProtKB. Some of these proteins have complete EC numbers that could not be mapped to KEGG pathways, and this is likely because those EC numbers have not been mapped to KEGG pathways. For many other proteins, the EC numbers are incomplete, and thus could not be inserted into specific reactions. Some of the most common of these EC numbers include EC:1.-.-.- (oxidoreductases), EC:2.1.1.- (methyltransferases), EC:2.3.1.- (acyltransferases transferring groups other than amino-acyl groups), EC:3.1.-.- (hydrolases acting on ester bonds) and EC:6.2.1.- (acid-thiol ligases).

3.5 Reordering of Matrix

The reordering of the matrix put all proteins in order of their pathways, allowing each pathway to be mapped and counted across all organisms.

3.6 Mapping Proteins to KEGG Pathway Diagrams

All metabolic pathway diagrams shown in KEGG for *M. tuberculosis* H37Rv were then used to map both pre-existing proteins (already characterised in KEGG) as well as those added through manual annotation onto reactions within the diagrams. This was completed for a total of 89 KEGG pathways shown to exist in H37Rv as well as some pathways not assigned to H37Rv but that might possibly actually exist. When proteins were mapped to pathways, some proteins identified using GO terms were then found not to be applicable to the assigned pathways; this was especially found true in cases

where GO terms were general terms without specific application to particular pathways.

3.7 Annotating Additional Reactions Using BLASTP

There is a well-established idea that sequence similarity suggests functional similarity (Koonin, 2005; Yu et al., 2004). It is therefore assumed that sequence similarity can be used to identify function. This study used BLASTP to match proteins between Mtb and closely related organisms in order to find additional functional homologs (no distinction was made between orthologs and paralogs) that could then be mapped to pathways. Since these BLASTP matches only looked at sequence similarity and did not take into account the origins of these matches (whether via orthology or paralogy), only closely related organisms were used. Using only closely related organisms reduces the chance for major evolutionary divergence, and thus corresponding functional divergence, between homologs. These matches enabled the annotation of additional reactions within the metabolome of Mtb.

In total, 1,217 H37Rv proteins had matches from the BLASTP runs, completing 288 reactions in 78 metabolic pathways. Some of these proteins were proteins already found in other reactions within the pathway and were simply annotated with additional functions. Others were matched multiple times with different proteins from related organisms, resulting in the assignment of multiple EC numbers to these proteins. Many other proteins were previously uncategorised into pathways but did have GO terms, EC numbers or gene names assigned. Lastly, 32 proteins were previously uncharacterised or putative but could be assigned to pathways based on these BLASTP matches.

3.7.1 Using BLASTP on Other Organisms

Of the 288 reactions completed using BLASTP matches, 136 were matches to proteins in *M. smegmatis* MC2 155. The remaining 152 reactions were completed with matches to proteins in eight other organisms. 18 reactions were completed with matches in *M. vanbaalenii*, 6 with *C. glutamicum* ATCC 13032 Kyowa Hakko, 20 with *N. farcinica*, 1 with *N. brasiliensis*, 81 with *R. jostii* RHA1, 4 with *R. erythropolis* PR4, 1 with *G. bronchialis* and 21 with *S. coelicolor.* When checking the BLASTP results, organisms were compared in the order written above because this is the order in which they are most closely related to Mtb. Thus, if a protein matched a protein in *M. vanbaalenii* then

that reaction would not be investigated for any of the other organisms. Since *M. vanbaalenii* is the most closely related of the proteins and the second best match with *M. tuberculosis* H37Rv after *M. smegmatis* MC2 155, the highest proportion of matches should thus be with this organism. The larger number of Mtb proteins matching proteins of *R. jostii* RHA1 could be due to the greater number of its proteins being characterised in KEGG; KEGG has more complete pathways for *R. jostii* RHA1 compared to the other organisms. This influences the results because all proteins used to identify BLASTP matches with Mtb were derived from the KEGG pathway diagrams for those organisms. However, this did not always prove true as other organisms such as *S. coelicolor* were also quite well characterised, while *N. farcinica* was usually less characterised but its proteins were matched quite often with proteins in Mtb. Thus, this result is likely caused by a combination of factors: the number of characterised reactions for each organism in KEGG and the phylogenetic relatedness of the organism to Mtb.

The BLASTP searches resulted in the elucidation of a number of additional proteins in KEGG pathways. A number of interesting pathways were found which could be completed or almost completed. On the other hand, some pathways already seemed to be complete, while others could not be added to at all. A few notable pathways include benzoate degradation, phenylalanine metabolism and glyoxylate and dicarboxylate metabolism. The results of these protein matches are shown in Table 3.4. Table 3.4 shows only those pathways that have been identified as 'interesting' for further discussion. The results for the remaining pathways are shown in Appendix A.

Additionally, three pathways with many newly annotated Mtb H37Rv proteins were investigated further by mapping the gene locations to the chromosome using Ensembl Bacteria. These pathways include 'benzoate degradation', 'phenylalanine metabolism' and 'glyoxylate and dicarboxylate metabolism'. 'Benzoate degradation' has four proteins that reside next to one another in two locations on the chromosome, 'phenylalanine metabolism' has five proteins from two reactions residing next to each other on the chromosome and 'glyoxylate and dicarboxylate metabolism' has two newly annotated proteins from the same pathway located next to each other. These will be discussed further in the following chapter.

Table 3.4 Results of annotations for proteins of *M. tuberculosis* H37Rv. The first column shows the KEGG metabolic pathway. The second column provides a count of all genes assigned to that pathway including those already categorised in KEGG and those added by the annotation methods described previously. The third column includes all proteins added through the annotation methods for each of the pathways (duplicates are not recorded). (EC) means that the protein was categorised into that pathway based on its EC number, (GO) means it was categorised based on its GO terms, and the rest are based on BLASTP matches. The organism in which the protein was matched is shown in parentheses: (Msmeg) = *M. smegmatis,* (Mvan) = *M. vanbaalenii,* (Cglut) = *C. glutamicum,* (Nfarc) = *N. farcinica,* (Nbra) = *N. brasiliensis,* (Rjost) = *R. jostii,* (Rery) = *R. erythropolis,* (Gbron) = *G. bronchialis* and (Strep) = *S. coelicolor.* Proteins marked with colour signify two types of proteins: blue means that the protein was previously uncategorised into any pathway at all, red means that that protein is uncharacterised, without any functional description in the databases and yellow means that a protein match was found for a strain other than *M. tuberculosis* H37Rv (the strain is shown in parentheses). The last column provides a count of all the likely pathway holes, or 'missing' (uncharacterised) reactions; sometimes the count of these missing reactions could not be determined. In these cases, two counts are provided based on different circumstances (existence or not existence of certain pathway branches). When no estimate could be made they were marked 'Unknown'.

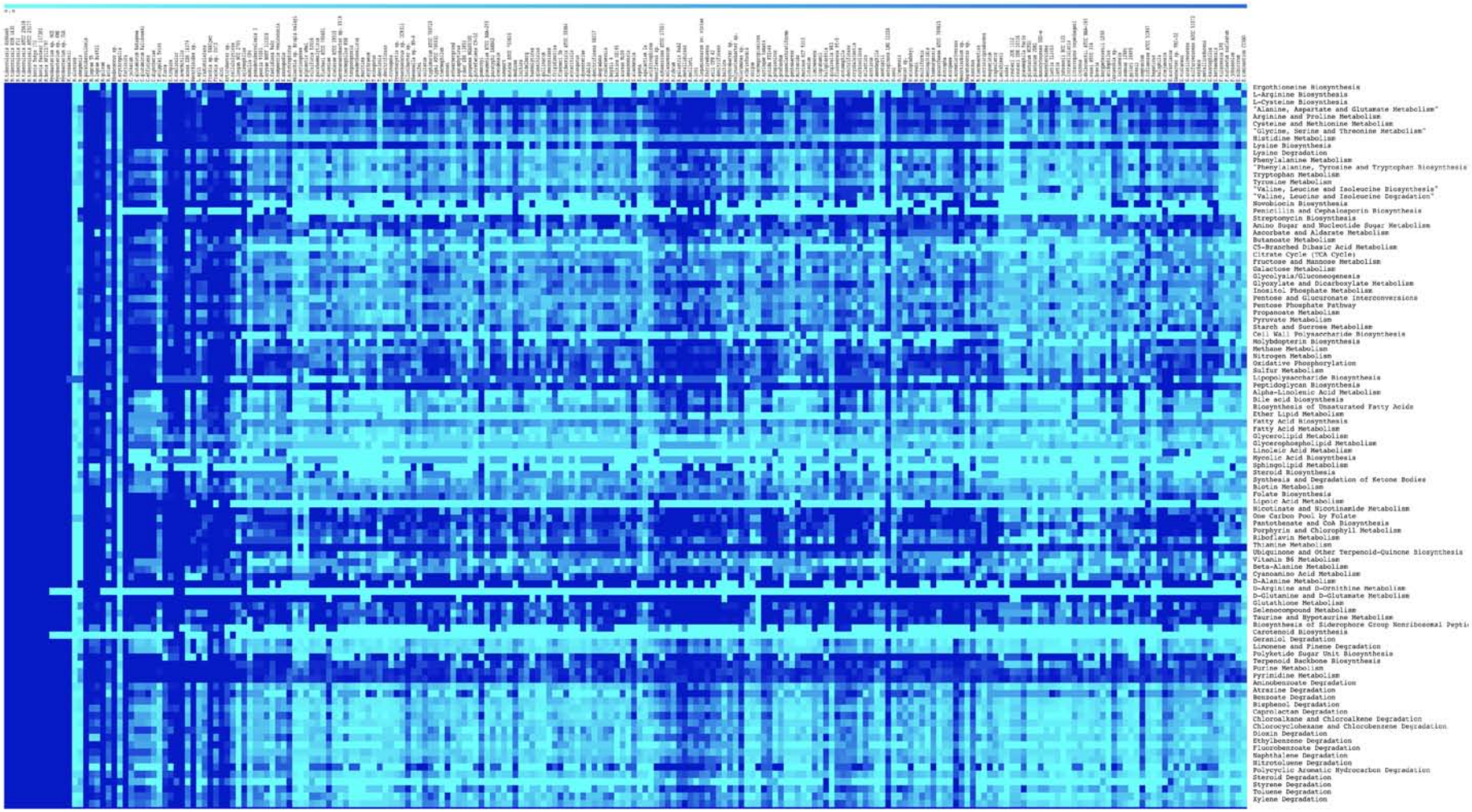
Pathway	# of Prot in Path	New Proteins Found	# Missing
Histidine Metabolism	41	Q50464 & O33219 & P96406 (EC); P0A678 (GO); O53494 (Strep)	3
Phenylalanine Metabolism	101	O53303 & O53533 & Q7DAC8 & P71818 & O53904 & P0A4X0 & P95153 & O07737 (Msmeg); P96261 & O53185 (Msmeg); Q7D5R0 & Q7D5Q9 (Msmeg); O06542 & O06541 & O53163 & P64016 & O53561 & P64014 & O53232 & Q7D9G0 & O53419 & P71540 & O53286 & O06414 & P64018 & O86369 & O07179 & P75019 & O50402 & P95279 & P96404 & P71621 & P71851 (Msmeg); O65936 (Msmeg); P96851 & P0A572 & O06420 & P64303 & O53622 & P64301 & O53327 (Msmeg); P95034 & P95146 (Rjost); O65936 (Rjost); O33340 & P96405 & P96417 & P63937 & P71823 & P71989 & O53816 & O50443 (Strep); P96843 & O06168 & O07411 & O53306 & O53406 & P95227 & O06417 & O05295 & P96396 & O07169 & O06831 & O53551 & O53521 & Q7D5D8 (Rjost); Q8VK36 (Rjost); O06168 (Rjost); O53157 (Rjost); P95162 (Nfarc); O07431 & O33219 (Cglut); P66781 & P95286 & O33292 & P69167 & P0A5Y4 & Q7D6M3 & P71824 & P95273 & O05919 & O33339 & P95033 & O33263 & O06544 & Q11150 (Mvan); P96847 & O53870 & P72039 & P0A678 (Rjost)	7
Arginine and Proline Metabolism	72	O69670 (GO); P64811 (GO); P96847 (GO); P0A4Y8 (GO)(repressor); O53258 (Msmeg); O05791 (Msmeg); P95041 (Msmeg); P95223 (Msmeg); Q11146 (Msmeg); O07760 (Msmeg) O53227 & Q7D7G7 (Msmeg); P0A5M4 (Nfarc); P0A508 & P95127 & P96890 & P71538 (Rjost); P63937 & O33340 & P96405 & P71823 & P96417 & P71989 & O53816 & Q7D5R7 & O50443 & P96824 (Rjost); P63498 & P96847 & P72039 & P0A678 (Rjost); P0A4X6 & O53379 & P63509 & P63504 & P63506 (Strep); O05578 & Q50708 (Nfarc)	19
Glycine, Serine and Threonine Metabolism	70	O53208 & P66801 & P71588 (GO); P71588 (GO); P66875 & P95199 & O53390 (Msmeg); P64263 & P95043 (Msmeg); P96405 & O33340 & P63937 & P71823 & P71989 & P96417 & O53816 & Q7D5R7 & P96824 & O50443 (Msmeg); Q8VKI0 (Msmeg); P63568 & P63504 & P63509 & O53379 & P0A4X6 & P71890 & P63506 & P71891 (Msmeg); Q50708 (Gbron); P71602 & P0A4X4 (Strep); O53533 & O07737 & O69693 & O53303 & O53904 & P0A4X0 & P71818 & P95153 & P72043 (Strep); O53272 (Strep); P71602 & P0A4X4 (Nbra)	8
Glycerolipid Metabolism	35	P0AX40 (EC); O53526 (EC); O53208 (EC); P0A4V2 & P0A4V4 & P0C5B9 & O53209 & P67210 & P96403 & O69707 & P71694 & P0A650 & O05879 & O06795 & O50400 & P67206 & P67204 & P67208 & O06343 & O53304 & C6DRC4 (EC); O07427 & O53321 & P95276 & O05805	1

		(Msmeg)	
Glyoxylate and Dicarboxylate Metabolism	72	O05790 (GO); P64118 & P65408 (GO); P95313 (Msmeg); Q8VK36 (Msmeg); P64016 & P96404 & P75019 & O53561 & O50402 & P71540 & P71621 & P64014 & O53286 & O53163 & O06542 & P71851 & P64018 & O53872 & O06414 & O06541 & O53232 & O06536 (Msmeg); P0A622 & P66946 & O53639 & O06335 & O53865 (Msmeg); P63935 & P71825 & Q79FJ2 & O06574 (Msmeg); P0A544 & O86322 (Msmeg); P96218 (Msmeg); O53929 (Rery); O65487 (Strep); P96405 & O33340 (Strep)	7
Fructose and Mannose Metabolism	73	A5TZ90 (H37Ra)(GO); A5WLC4 (F11)(EC); O07737 & O53303 & O53533 & P95153 & P0A4X0 & O69693 & O53904 & O53146 & Q7DAC8 & P72043 & O07721 (Msmeg); P96825 & P71853 & O53547 & P66781 & P71824 & P95033 & O33292 & O53863 & P96841 & O53398 & O50417 & P0A5Y4 & O53665 & O53302 & P69167 & O50460 & P95101 & P66777 & P95273 & P95286 & P71821 & O33263 & O33339 & O05919 & Q7D6M3 & P71871 & P71852 & O53927 & P66779 & O06348 & P95150 & O06413 & P95185 & O05842 & O53613 & O53726 & O53537 & O07230 & O53324 & Q11150 & Q10782 & P96202 (Msmeg); O07248 (Msmeg)	5 (or 15)
Butanoate Metabolism	132	P95286 & P69167 & P71824 & P96841 & P66781 & P95273 & O33339 & O53863 & P71871 & P0A5Y4 & P66777 & O53398 & O05919 & O53665 & P95101 & Q7D6M3 & P71853 & O06413 & O33292 & O50460 & P95033 & O06348 & O53302 & P95150 & O53547 (Msmeg); P95313 (Msmeg); P96825 & O50417 & P71821 & O33263 & P71852 & O53927 & P66779 & O07737 & O53303 & O53533 & O69693 & P71818 & Q7DAC8 & O53904 & P95153 & O53146 & P0A4X0 & P72043 & P95185 & O05842 & O53613 & O53726 & O53537 & O07230 & O53324 & Q11150 & Q10782 & P96202 (Msmeg); P64016 & P96404 & P75019 & O53561 & O50402 & P71540 & P71621 & P64014 & O53286 & O53163 & O06542 & P71851 & P64018 & O53872 & O06414 & O06541 & O53232 & O06536 (Msmeg); O69635 & P71716 & P95227 & Q7D5D8 & O06831 & O06168 & O07411 & O53406 & Q10878 & P96396 & P71717 & O53521 & P96283 (Msmeg); O06334 & P71867 (Msmeg); P71850 (Nfarc); P63427 & P96397 & O06164 & O53815 & O33229 & P95208 & P63429 & O33331 & O86319 & P96831 & P71539 & P95208 & O53549 & P95186 & P96808 & O53577 & P96844 & P71858 & P96842 & P95228 & O53926 (Rjost); O33185 (Nfarc); O53146 & P72043 & P95185 (Strep)	5
Sulfur Metabolism	59	A5U3C0 (EC); P71615 (EC); P0A534 (EC); O69722 & P0A4W2 & O50454 & O53482 & O69724 & O53899 & P71620 & O53618 & O53149 & O05779 & P63401 & P95155 & O65934 & O69631 & P63393 & P63399 & P63357 & P63395 (Msmeg); P96809 & P96253 & P64761 & P71844 (Msmeg); P64769 & P95079 & P95159 & Q10814 (Msmeg); O07190 & O53832 & P96205 & O33189 & P95302 & O86311 & O53164 (Msmeg)	1
Geraniol Degradation	73	O05842 & P95273 & P66781 & O33263 & O53398 & P71853 & P71824 & P95286 & P96825 & O33292 & P95033 & P71821 & P71871 & O33339 & P69167 & P66777 & O06413 (Msmeg); O06334 & P71867 (Msmeg); O53567 (EC); O86319 & O33229 (Nbra)	6
Limonene and Pinene Degradation	104	L0T647 (EC); P96825 & P71853 & O53547 & P66781 & P71824 & P95033 & O33292 & O53863 & P96841 & O53398 & O50417 & P0A5Y4 & O53665 & O53302 & P69167 & O50460 & P95101 & P66777 & P95273 & P95286 & P71821 & O33263 & O33339 & O05919 & Q7D6M3 & P71871 & P71852 & O53927 & P66779 & O06348 & P95150 & O06413 & O07737 & O53303 & O53533 & O69693 & P71818 & Q7DAC8 & O53904 & P95153 & O53146 & P0A4X0 & P72043 & P95185 & O05842 &	10

		O53613 & O53726 & O53537 & O07230 & O53324 & Q11150 & Q10782 & P96202 (Msmeg)	
Porphyrin and Chlorophyll Metabolism	35	P64803 (EC); O69680 (EC); P64955 (GO); P66877 (Rjost); P71751 & P95216 (Strep)	3
Biotin Metabolism	38	P0A5Y4 (EC); P63456 & P63454 & O53579 & Q10977 & O86335 & P96284 & P94996 & L0TA10 & O53901 & P96202 & P96291 & L0T5X5 & P71718 & O06586 & O65933 & P96204 & O07798 (Msmeg); P0A5Y6 & P71871 & O05919 & P71824 (Strep)	5
One Carbon Pool by Folate	16	C6DVW3 (GO)(KZN); O53217 (EC); O05575 (Msmeg)	3
Folate Biosynthesis	21	P62589 (EC)	1
Benzoate Degradation	123	O53567 (EC); P96825 & P71853 & O53547 & P66781 & P71824 & P95033 & O33292 & O53863 & P96841 & O53398 & O50417 & P0A5Y4 & O53665 & O53302 & P69167 & O50460 & P95101 & P66777 & P95273 & P95286 & P71821 & O33263 & O33339 & O05919 (Msmeg); P0A666 (Msmeg); P95277 & P71871 & Q7ARS9 & Q11150 & P96853 & Q7D6M3 & P71846 & O06413 (Msmeg); O07243 (Msmeg); P0A572 & O06420 & P66777 (dup) & O53321 & O53327 & P96935 (Msmeg); P71716 & O06168 & O53306 & O53406 & O07411 & P95227 & P96843 & O06831 & O05295 & P96396 & O07169 & P96283 & O69635 & O06417 & Q10878 (Msmeg); P71852 & O06348 & O06544 (Msmeg); P71832 (Rjost); P95034 & P95146 (Mvan); P96405 & P63937 & P71823 & O33340 & P96417 & P71989 & O53816 & Q7D5R7 & P96824 & O50443 (Nfarc); O65936 (Rjost); O53927 & O50417 & O50460 (Rjost); P65425 (Rjost); P96850 & O33319 & O86347 (Rjost); O53555(Rjost); P95118 (Rjost); P65083 (Rjost)	6 (or 12)
Xylene Degradation	78	O53303 & O53533 & Q7DAC8 & P71818 & O53904 & P0A4X0 & P95153 & O07737 (Msmeg); P96405 & P71823 & O33340 & P96417 & P71989 & P63937 & O53816 & Q7D5R7 & P96824 & O50443 (Msmeg); P95277 & P66781 & P71871 & O33339 & P71824 & P96841 & Q7ARS9 & P69167 & O33292 & P95286 & O05919 & Q11150 & O53398 & P95033 & P96853 & P95273 & O33263 & Q7D6M3 & P71846 & O53863 & P0A5Y4 & P95101 & O06413 & P71853 & P71821 & O86347 & O53311 (Msmeg); P96850 & O33319 (Rjost); O53665 & P71852 & O53927 & P96825 & O50417 & O50460 & O53302 (Rjost); P96851 & P0A572 & P64303 & O06266 & O53327 & P95276 & P64301 & P96811 & O69638 & O06420 & O53321 & O86348 (Rjost); P95034 & P95146 & O53674 & O05875 & Q79FW1 & O53355 (Rjost); P95034 & P95146 (Mvan)	2
Caprolactam Degradation	68	P0A4X0 (GO)(confirmed with Cglut); O53303 & O53533 & Q7DAC8 & P71818 & O53904 & P0A4X0 & P95153 & O07737 (Msmeg); P64745 & P71662 & P96223 & O53762 & O53300 & P64765 & O53294 (Msmeg); P96405 & O53816 & P63937 (Msmeg); P96397 & O53815 & P63427 & O33229 & O06164 & P95208 & P63429 & O33331 & P96808 & O86319 & P71539 & P96831 & O53549 & P95280 & P95187 & P95186 & P71858 & P95281 & P96842 & O53577 (Msmeg)	2
Polycyclic Aromatic Hydrocarbon Degradation	73	P96850 (Msmeg); P63945 (Mvan); P96405 & O33340 & P71989 & P71823 & P63937 & P96417 & O53816 & Q7D5R7 & O50443 & P96824 (Mvan); P63945 (Mvan); O53772 & Q11058 (Rjost); O05301 & O06339 (Rjost); O53311 & P95034 & P95146 & O53674 & P66006 & O07927 & O06598 & O53641 & P96839 (Rjost); O86347 & P96853 & O05875 & P71846 & O33319 (Rjost)	7

3.8 Analysing the Phylogenetic Profile

The phylogenetic profile originally contained 1,161 organisms with ortholog data from 99 different metabolic pathways. After removing species with orthologs in less than 50 pathways (showing incomplete data) and removing multiple strains of the same species so that only one entry for each species remained (to reduce the image sizes), the matrix contained 371 organisms. The number of orthologs for each species in each pathway was then divided by the total number of proteins for that pathway in *M. tuberculosis* H37Rv, resulting in the frequency of orthologous proteins for each species in each pathway. The resulting data was assembled into a heat map, shown in Figure 3.1 (pages 51-52). The average frequency across all species (except for organisms showing incomplete data) for each pathway was also calculated in order to compare pathways. These averages are shown in Figure 3.2.


Ergothioneine Biosynthesis
L-Arginine Biosynthesis
L-Cysteine Biosynthesis
"Alanine, Aspartate and Glutamate Metabolism"
Arginine and Proline Metabolism
Cysteine and Methionine Metabolism
"Glycine, Serine and Threonine Metabolism"
Histidine Metabolism
Lysine Biosynthesis
Lysine Degradation
Phenylalanine Metabolism
"Phenylalanine, Tyrosine and Tryptophan Biosynthesis
Tryptophan Metabolism
Tyrosine Metabolism
"Valine, Leucine and Isoleucine Biosynthesis"
"Valine, Leucine and Isoleucine Degradation"
Novobiocin Biosynthesis
Penicillin and Cephalosporin Biosynthesis
Streptomycin Biosynthesis
Amino Sugar and Nucleotide Sugar Metabolism
Ascorbate and Aldarate Metabolism
Butanoate Metabolism
C5-Branched Dibasic Acid Metabolism
Citrate Cycle (TCA Cycle)
Fructose and Mannose Metabolism
Galactose Metabolism
Glycolysis/Gluconeogenesis
Glyoxylate and Dicarboxylate Metabolism
Inositol Phosphate Metabolism
Pentose and Glucuronate Interconversions
Pentose Phosphate Pathway
Propanoate Metabolism
Pyruvate Metabolism
Starch and Sucrose Metabolism
Cell Wall Polysaccharide Biosynthesis
Molybdopterin Biosynthesis
Methane Metabolism
Nitrogen Metabolism
Oxidative Phosphorylation
Sulfur Metabolism
Lipopolysaccharide Biosynthesis
Peptidoglycan Biosynthesis
Alpha-Linolenic Acid Metabolism
Bile acid biosynthesis
Biosynthesis of Unsaturated Fatty Acids
Ether Lipid Metabolism
Fatty Acid Biosynthesis
Fatty Acid Metabolism
Glycerolipid Metabolism
Glycerophospholipid Metabolism
Linoleic Acid Metabolism
Mycolic Acid Biosynthesis
Sphingolipid Metabolism
Steroid Biosynthesis
Synthesis and Degradation of Ketone Bodies
Biotin Metabolism
Folate Biosynthesis
Lipoic Acid Metabolism
Nicotinate and Nicotinamide Metabolism
One Carbon Pool by Folate
Pantothenate and CoA Biosynthesis
Porphyrin and Chlorophyll Metabolism
Riboflavin Metabolism
Thiamine Metabolism
Ubiquinone and Other Terpenoid-Quinone Biosynthesis
Vitamin B6 Metabolism
Beta-Alanine Metabolism
Cyanoamino Acid Metabolism
D-Alanine Metabolism
D-Arginine and D-Ornithine Metabolism
D-Glutamine and D-Glutamate Metabolism
Glutathione Metabolism
Selenocompound Metabolism
Taurine and Hypotaurine Metabolism
Biosynthesis of Siderophore Group Nonribosomal Pepti
Carotenoid Biosynthesis
Geraniol Degradation
Limonene and Pinene Degradation
Polyketide Sugar Unit Biosynthesis
Terpenoid Backbone Biosynthesis
Purine Metabolism
Pyrimidine Metabolism
Aminobenzoate Degradation
Atrazine Degradation
Benzoate Degradation
Bisphenol Degradation
Caprolactam Degradation
Chloroalkane and Chloroalkene Degradation
Chlorocyclohexane and Chlorobenzene Degradation
Dioxin Degradation
Ethylbenzene Degradation
Fluorobenzoate Degradation
Naphthalene Degradation
Nitrotoluene Degradation
Polycyclic Aromatic Hydrocarbon Degradation
Steroid Degradation
Styrene Degradation
Toluene Degradation
Xylene Degradation

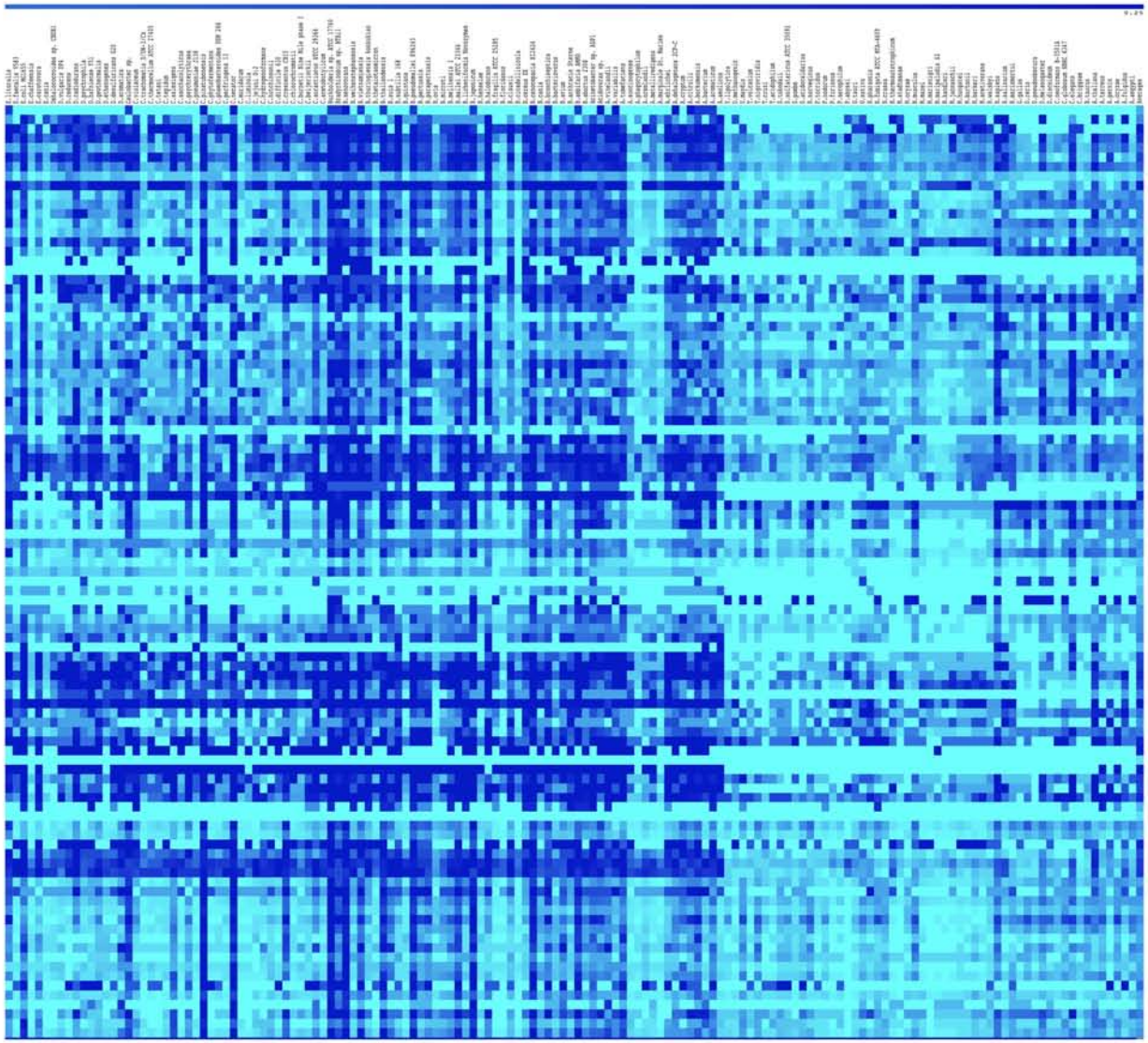


Figure 3.1: This image shows a heat map of the frequencies of proteins per pathway in each organism by dividing the total count of the homologous proteins for each organism by the total proteins belonging to that pathway in M. tuberculosis H37Rv. Thus, the coloured boxes are displayed on a scale of 0.0-1.0, lightest to darkest. A list of the organisms shown is included in Appendix B because they are not legible on these images.

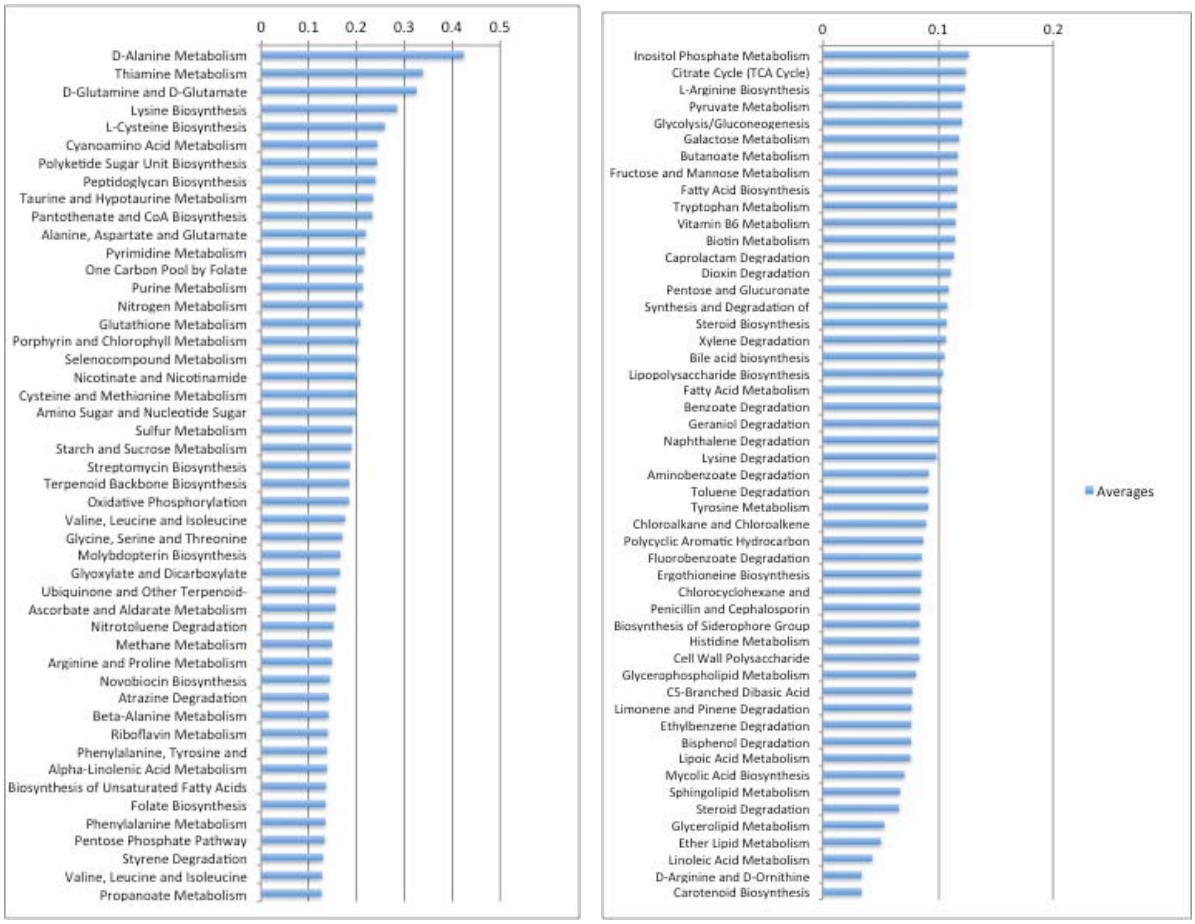
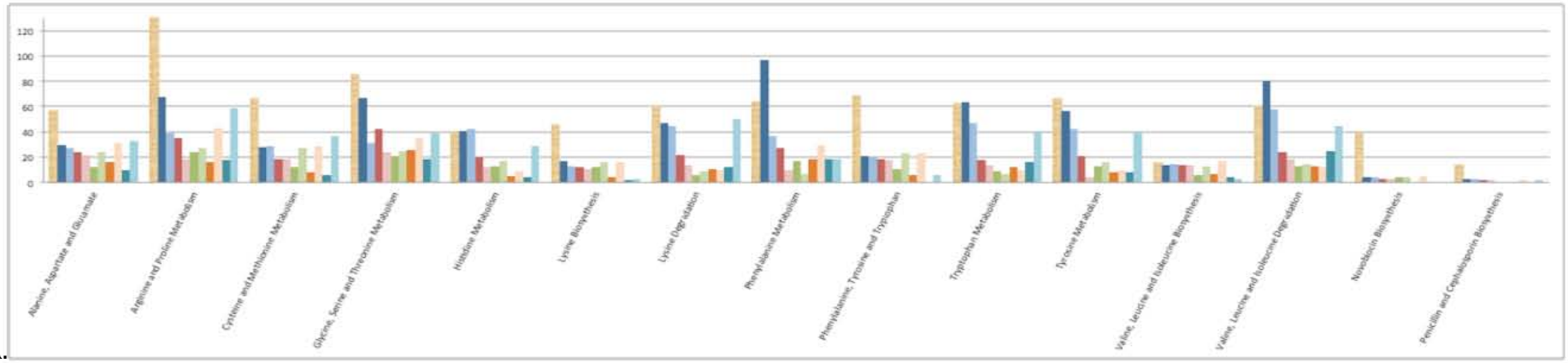


Figure 3.2 These charts show the average frequency of homologs for all species in the phylogenetic profile according to pathway. Those with the highest averages show pathways in which relatively higher numbers of orthologs were found across all species within the phylogenetic profile.

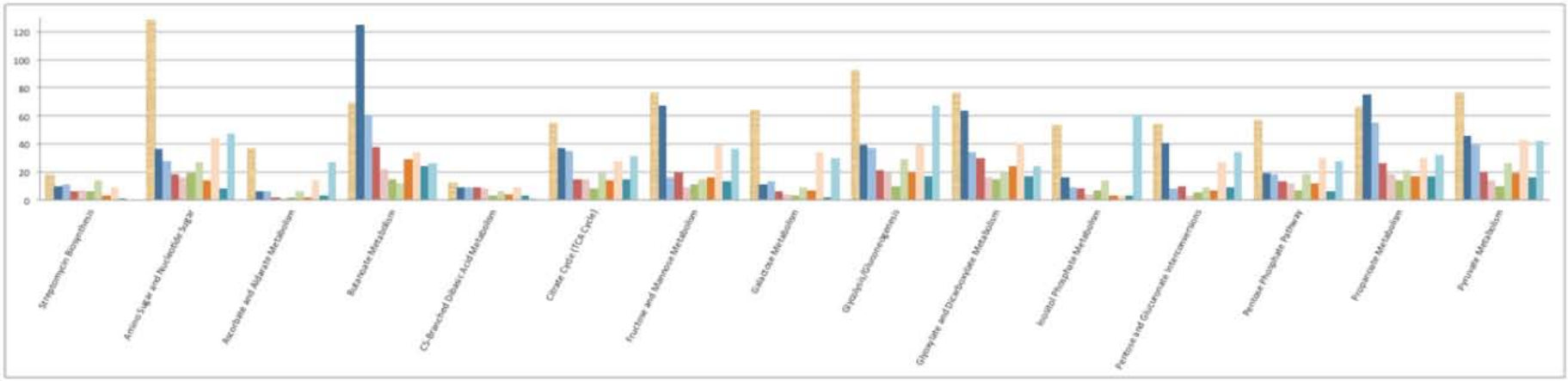
3.9 Deriving the Pathway Summary Totals for *Escherichia coli*

The pathway summary totals for *E. coli*, *M. leprae* TN, *C. glutamicum* and *H. sapiens* were derived through KEGG (already characterised reactions in KEGG) and compared to those found in *M. tuberculosis* H37Rv as well as those marked as orthologs in the phylogenetic profile. A graph showing the differences between the number of proteins in the reference pathway, Mtb, *M. leprae, C. glutamicum, E. coli* and *H. sapiens* is shown in Figure 3.3 (pages 54-55). This graph also shows the differences between the numbers of proteins shown in the created phylogenetic profile, which has many added new proteins, versus the number of proteins for each of the five organisms in each pathway in KEGG.

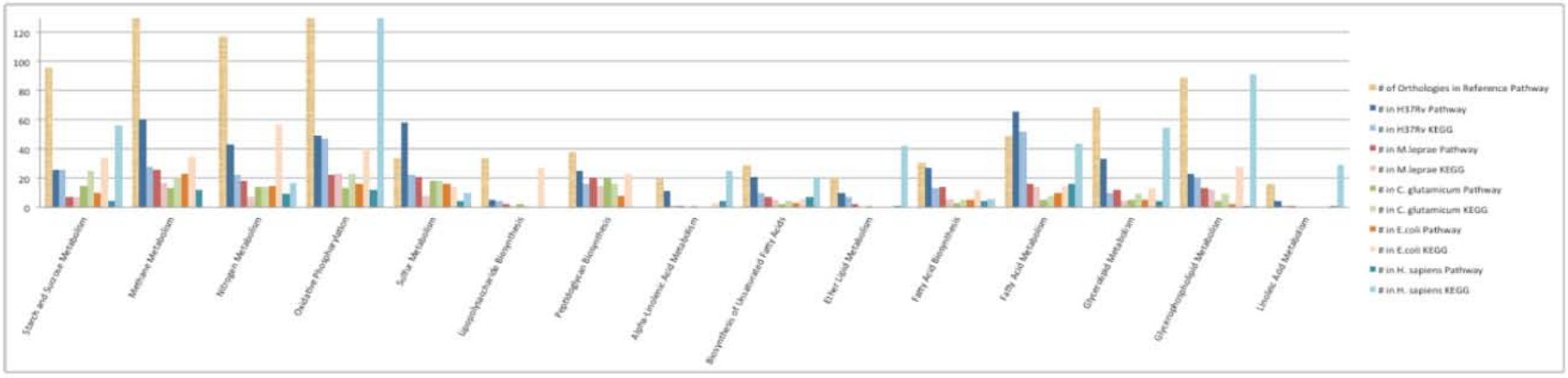
A.


120
100
80
60
40
20
0
Alanine, Aspartate and Glutamate
Arginine and Proline Metabolism
Cysteine and Methionine Metabolism
Glycine, Serine and Threonine Metabolism
Histidine Metabolism
Lysine Biosynthesis
Lysine Degradation
Phenylalanine Metabolism
Phenylalanine, Tyrosine and Tryptophan
Tryptophan Metabolism
Tyrosine Metabolism
Valine, Leucine and Isoleucine Biosynthesis
Valine, Leucine and Isoleucine Degradation
Novobiocin Biosynthesis
Penicillin and Cephalosporin Biosynthesis

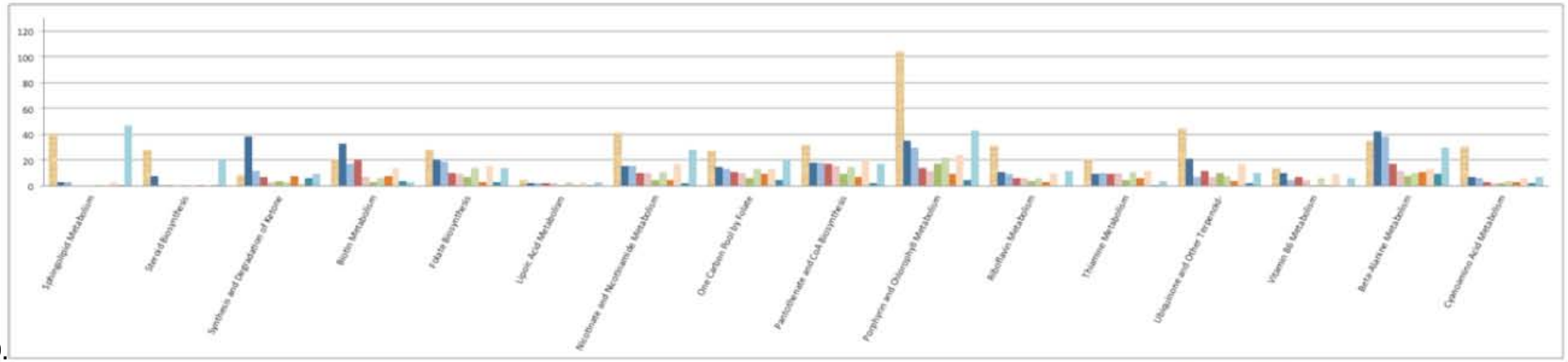
B.


120
100
80
60
40
20
0
Streptomycin Biosynthesis
Amino Sugar and Nucleotide Sugar
Ascorbate and Aldarate Metabolism
Butanoate Metabolism
C5-Branched Dibasic Acid Metabolism
Citrate Cycle (TCA Cycle)
Fructose and Mannose Metabolism
Galactose Metabolism
Glycolysis/Gluconeogenesis
Glyoxylate and Dicarboxylate Metabolism
Inositol Phosphate Metabolism
Pentose and Glucuronate Interconversions
Pentose Phosphate Pathway
Propanoate Metabolism
Pyruvate Metabolism

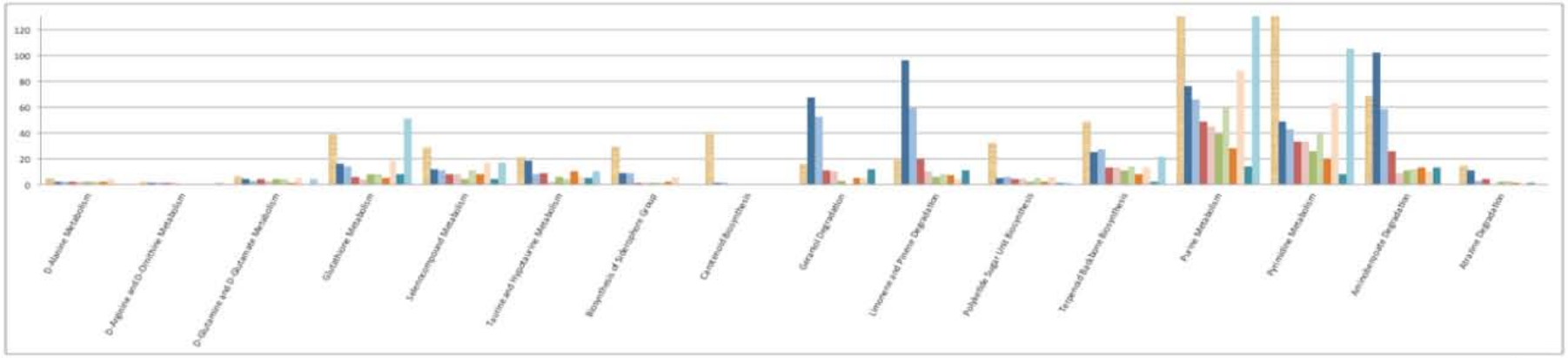
C.


120
100
80
60
40
20
0
Starch and Sucrose Metabolism
Methane Metabolism
Nitrogen Metabolism
Oxidative Phosphorylation
Sulfur Metabolism
Lipopolysaccharide Biosynthesis
Peptidoglycan Biosynthesis
Alpha-Linolenic Acid Metabolism
Biosynthesis of Unsaturated Fatty Acids
Ether Lipid Metabolism
Fatty Acid Biosynthesis
Fatty Acid Metabolism
Glycerolipid Metabolism
Glycerophospholipid Metabolism
Linoleic Acid Metabolism
# of Orthologies in Reference Pathway
in H37Rv Pathway
in H37Rv KEGG
in M.leprae Pathway
in M.leprae KEGG
in C. glutamicum Pathway
in C. glutamicum KEGG
in E.coli Pathway
in E.coli KEGG
in H. sapiens Pathway
in H. sapiens KEGG

D.



E.



F.

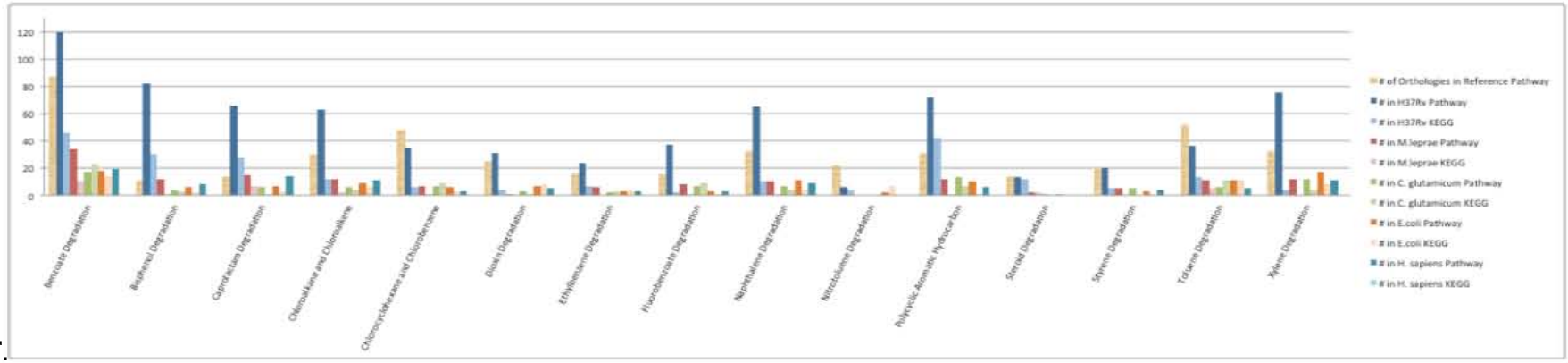


Figure 3.3 Bar graphs (A-F) comparing the numbers of proteins per pathway for KEGG reference pathways, *M. tuberculosis* H37Rv, *M. leprae* TN, *C. glutamicum*, *E. coli* and *H. sapiens*. Two counts for each organism are made: one from the matrix compiled in this research (labelled Pathway in the key) and the second from the KEGG orthology proteins for each organism (labelled KEGG in the key).

Using these graphs notable pathways were identified with at least one of five characteristics: those for which Mtb and *M. leprae* share both many and very few orthologous proteins, those for which the pathway totals are very different between *M. tuberculosis* H37Rv and *E. coli*, those for which Mtb and *C. glutamicum*, as a facultative anaerobe, share many orthologs, and those in which Mtb and *H. sapiens* do not share many orthologs. In many cases, the totals found for *M. tuberculosis* H37Rv are much higher than *E. coli* and the other organisms. This is likely due to the acceptance of all BLASTP matches with an e-value lower than 10^{-6}; some matches could simply be closely related proteins that function in other similar pathways. While this could bias the results towards higher counts for proteins in pathways of *M. tuberculosis* H37Rv, in some cases the results for *E. coli* or other organisms were substantially higher than those of *M. tuberculosis* H37Rv. Pathways that fulfil one or more of the five aforementioned characteristics will be discussed in the following chapter.

3.10 Identifying 'Missing' Pathways or Pathway Holes

Using all methods previously described to identify 'missing' pathways or pathway holes, meaning essential reactions for which no gene has been identified, a total of 363 possible pathway holes were identified (shown in Appendix C). Pathway holes have been identified for almost all metabolic pathways, and particular notable pathway holes are discussed in the following chapter. Only one pathway hole has been identified as missing in multiple pathways, the rest of the 361 pathway holes are all 'unique' reactions. These pathway holes might occur in instances where Mtb utilises the host metabolome in order to accomplish certain metabolic needs. Alternatively, enzymes might simply not yet be identified for these reactions in the genome of Mtb.

3.10.1 Pathway Tools

The Pathway Tools "pathway holes" file contained 198 missing reactions for which no corresponding enzyme had been identified for these metabolic pathways in *M. tuberculosis* H37Rv by the Pathway Tools software. 100 of these pathway holes were in fact filled by enzymes in KEGG and identified using the annotation methods described above, resulting in only 98 pathway holes for which no corresponding enzyme has been identified in *M. tuberculosis* H37Rv.

3.10.2 KEGG

The KEGG global map for *M. tuberculosis* H37Rv was checked for possible missing pathways. Certain reactions on this global map are grey in colour and others show colours, either pale or bright. Bright colours show instances in which enzymes have been identified for reactions in *M. tuberculosis* H37Rv. It is believed that pale colours show reactions which are thought to exist but are uncharacterised by enzymes. Grey reactions are those that are thought not to exist in *M. tuberculosis* H37Rv. Thus, reactions that were shown in pale colours when the pathways were highlighted are marked as possible missing reactions. This pale colour was displayed for 125 of the possible missing reactions. This information is used in conjunction with other evidence to determine pathway holes for Mtb.

3.10.3 Existence in Closely Related Organisms

Of all the identified possible pathway hole reactions, 61 had corresponding enzymes identified in *M. smegmatis*. As *M. smegmatis* is a closely related organism within the Mycobacteria, it is possible that these uncharacterised pathways are in fact 'missing' or holes in Mtb. This is also assumed true for characterised pathways in the other closely related organisms.

3.10.4 Visual Evidence for Pathway Holes

For 112 reactions, none of the first three instances were fulfilled but yet reactions were still classified as 'missing' based on observational evidence. These occur in cases where an uncharacterised reaction might be surrounded by other reactions with identified enzymes. For example, in pathway chains or branches in which enzymes characterised multiple reactions but yet some were uncharacterised it would seem most likely that these uncharacterised reactions were in fact missing.

3.11 Conclusion

Thus, a substantial number of Mtb H37Rv proteins have been annotated using the methods described in this paper. Additionally, these annotations have been used to compare the metabolic map of Mtb across the phylogenetic profile and identify pathway holes, or functional gaps that remain to be completed either by Mtb or the host.

4 Discussion

The metabolic network characterisation efforts and creation of the phylogenetic profile elucidates a number of interesting aspects, and these will be discussed in this chapter. First, an evaluation of the methods used to annotate pathways as well as determine important pathways reveals some strengths and weaknesses of this research. These are discussed below, followed by an examination of the interesting results of these various methods. Firstly, the manual annotation plus the BLASTP results from multiple organisms show a number of interesting pathways that could be filled and/or possibly completed. Secondly, the comparison of total proteins per pathway between organisms shows an additional number of interesting pathways, which will be described below. Lastly, the missing pathway chart shows even more interesting pathways in terms of the reactions still classified as missing after using several methods to characterise pathways.

4.1 Assessment of Methods Used

The methods used to add annotations to the *M. tuberculosis* H37Rv genome and compare the organisms within the phylogenetic profile provide a number of interesting results. However, some of these methods also cause potential confounding factors in terms of the interpretation of results.

One of the first potential factors affecting this work is the choice of KEGG as the main pathway database (Ogata et al., 1999). The choice of pathway database can have a huge effect on the results for a number of reasons. Firstly, pathway description and naming varies widely across databases and literature. Many pathways overlap and this causes discrepancies as well as repetition within databases. Additionally, some reactions marked as missing may in fact only be artefacts of other pathways. For example, if reactions are replicated within other pathways, they may show up in a pathway or a branch of a pathway, even if that section of a pathway doesn't exist within the organism. This may also cause problems in cases where compounds or metabolites are used in other pathways. For example, reactions may produce a particular metabolite within a pathway, with all following reactions showing as non-annotated; it could be assumed that the subsequent reactions might be missing in the organism. However, this

compound may be used in another pathway, and might not prove to be actually missing. Unlike KEGG, MetaCyc unlinks commonly repeated reactions, thus reducing the redundancy in the database. KEGG pathway maps include all known reactions related to a particular topic or goal, regardless of the species or kingdoms in which they occur. Since many pathway variants occur, when different organisms use different reactions or pathways in order to achieve the same metabolic goal, many reactions in KEGG diagrams are specific for unrelated organisms, such as perhaps mammals, and do not exist in Mtb (Caspi et al., 2013). The main annotations and pathways in this research were extracted from the KEGG database, because the diagrams in KEGG were clearer and more manageable. Since many of these KEGG diagrams were quite large, some were broken into pieces of pathways derived from the MetaCyc database. Since functional annotations were obtained from KEGG, there could be false negatives in that some MetaCyc reactions are not included in the KEGG database. Additionally, and probably having a stronger effect on the results, there could be false positives for characterised and missing reactions due to the overlap between pathways. Reactions that were characterised or found to be missing could rather be artefacts of other pathways. This could affect certain pathways by enabling them to be characterised and deemed functional in *M. tuberculosis* H37Rv when in fact they are only present because the reactions are shared with other pathways. Thus the obtained results might require further experimentation in order to confirm or refute them.

A second potential issue is the use of primary sequence structure to determine homology and EC number to determine functional characterisation. Protein matches were made to annotate the genome of *M. tuberculosis* H37Rv by using BLASTP to find protein matches for already annotated enzymes in KEGG. Studies have found that proteins in pathway alignments (aligning pathways from two or more organisms to find proteins) did not necessarily pair with the best sequence match of the other pathway. This suggests that if a multifunctional protein of one organism experiences duplication and specialisation in another organism, then it could have different orthologs in different metabolic pathways (Kelley et al., 2003). This could have an effect on the research conducted here, since additional reactions were annotated based on sequence similarity and not on protein-protein interactions. In addition to sequence similarity, EC number assignments were used to characterise some proteins. A previous study

aligning the metabolic pathways of *Escherichia coli* and *Saccharomyces cerevisiae* used EC number rather than sequence similarity to form alignments; this allowed the authors to locate functional orthologs resulting from both convergent evolution, in which non-homologous proteins evolved the same function, or divergent evolution, in which orthologous proteins diverge in sequence while retaining the same function. Thus they showed the efficacy of using functional classification in addition to sequence-based classification in order to compare pathways (Pinter et al., 2005). In the current study, functional classification (for example EC numbers and GO terms) has been used in addition to sequence-similarity in order to complete functional gaps in the metabolism of Mtb. This also provides a basis for the technique of using a protein's EC number to assign that protein to a pathway or reaction. However, this means that these functional classifications rely on the accuracy of the EC numbers and GO terms.

The phylogenetic comparison itself can be affected by the use of the KEGG database. It is hard to compare KEGG maps phylogenetically because of the fact that they encompass all reactions related to a particular topic; if one organism has less coverage of a pathway it could be because of different branches within the pathway. Alternatively, two organisms could show the same number of reactions for a pathway, but the actual mapping shows entirely different metabolisms, with the two organisms utilising completely different portions of the map with no or little overlap (Altman et al., 2013). While this means that pathway comparisons are more uncertain and less predictive in KEGG, the larger structure of KEGG maps makes it easier to compare all pathways of the metabolome as a whole. In this study, this factor should not be a major problem in comparing pathway totals of organisms to *M. tuberculosis* H37Rv since these totals were calculated using only orthologous proteins of *M. tuberculosis* H37Rv. On the other hand, organisms (besides *M. tuberculosis* H37Rv) shown to have the same number of proteins in pathways might have different parts of the pathways and thus cannot be compared to one another. Additionally, the protein counts of proteins derived from the KEGG orthology lists include duplicates of proteins; when a protein catalyses multiple reactions within one pathway, it is counted twice in these summary totals. This will sometimes lead to the KEGG protein counts being greater than those summarised from the matrix, and so this must be taken into account when evaluating these results.

The prediction of missing proteins can also be effected by the used of BLASTP to find protein matches with proteins from closely related organisms and characterise pathways in *M. tuberculosis* H37Rv. In previous research, well-characterised pathways of model organisms have been used to predict homologous pathways in other prokaryotes (Mao et al., 2012). Pathway holes are missing pieces in a mapped pathway, and can result from two conditions when performing pathway prediction in this manner. Firstly, genes of the template pathways may not always map to any gene in the organism. Secondly, the organism may possess reactions or pathways that the template pathways do not have. Thus, some 'missing' reactions may be false positives while other reactions not classified as 'missing' may prove to be false negatives.

4.2 Interesting Pathways Filled in *M. tuberculosis*

These pathways are interesting due to their completion where they previously had gaps or because of their functions within the genome. Benzoate degradation was chosen because it belongs to the xenobiotics biodegradation and metabolism category, to which many pathways have been added and partially completed. Xenobiotic biodegradation and metabolism is also particularly interesting because many of the reactions comprising this category do not exist in MetaCyc, the other major pathway database (Altman et al., 2013). Next phenylalanine metabolism was chosen from the amino acid metabolism category, because the major amino acid pathways are highly conserved among organisms, and it has been suggested that these pathways evolved prior to the divergence of organisms into the kingdoms of Archaea, Bacteria and Eukarya (Hochuli et al., 1999; Pinter et al., 2005). The last pathway examined, glyoxylate and dicarboxylate metabolism, belongs to the carbohydrate metabolism category, chosen because of its necessity for survival and essentiality for growth.

4.2.1 Benzoate Degradation

Aromatic compounds are incredibly widely distributed in nature and comprise one-quarter of the biomass on Earth (Valderrama et al., 2012). The degradation of aromatic compounds can provide a source of nutrients and energy (Caspi et al., 2012). One of the most distinctive steps in the degradation of aromatic compounds is ring cleavage although not all aromatic compound modifications cause fission (Evans, 1963). Degradation of aromatic compounds can occur in both aerobic and anaerobic

conditions via four pathways including aerobic, hybrid, anaerobic and strict anaerobic metabolism pathways. Hybrid metabolism occurs in aerobic conditions but uses similar reactive steps as anaerobic metabolism, while strict anaerobic metabolism uses a different reduction reaction that is not ATP-dependent, compared to typical anaerobic metabolism (Fuchs, 2008). However, these two pathways do not appear in the KEGG diagram for benzoate degradation and thus will not be discussed further.

Aerobic degradation of aromatic compounds requires extensive use of molecular oxygen and a huge number of oxygenases have evolved to act on aromatic compounds (Harwood et al., 1998). The initial goal of aerobic aromatic metabolism is to remove constituent groups and replace them with hydroxyl groups, thus hydroxylating the aromatic compound into central intermediates. Consequently, the central intermediates of aerobic aromatic degradation are easily attacked via oxidation, and a dioxygenase then cleaves the ring by incorporating two oxygen atoms from O_2 (Heider and Fuchs, 1997).

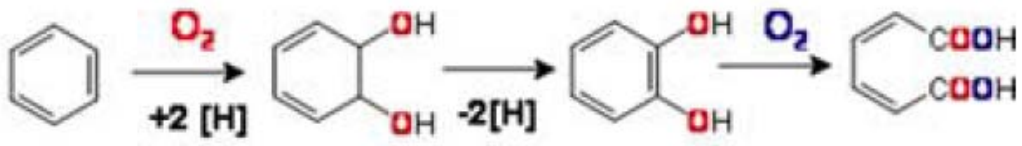


Figure 4.1 This image shows the general aerobic metabolism of aromatic compounds. The first reaction shown is the hydroxylating step in which two hydroxyl groups are added to the ring. Next two hydrogen atoms are removed to reform the benzene ring. The last reaction shows the ring cleavage step in which a dioxygenase utilises O2 to cleave the benzene ring (Fuchs, 2008).

Aerobic degradation of benzoate relies on oxygen as an essential co-substrate of the oxygenases that hydroxylate, or activate, and cleave, or dearomatise, the benzoate. Aerobic degradation is the classical form by which bacteria degrade benzoate and relies on the intermediary compound known as catechol (Valderrama et al., 2012). An alternative pathway to metabolise benzoate relies on the intermediary compound known as protocatechuate, or 3,4-dihydroxybenzoate (Fuchs, 2008). Two pathways are available to degrade catechol, 'catechol degradation II' and 'catechol degradation to β-ketoadipate' (Williams and Murray, 1974). Together, these two pathways account for the degradation of benzoate via catechol. Aerobic degradation can also occur via the intermediate protocatechuate using parallel reactions with 'catechol degradation to β-ketoadipate' which then converge at '3-oxoadipate degradation'. These two aerobic

pathways are widely distributed in soil bacteria and have been shown to be highly conserved in diverse bacteria (Harwood and Parales, 1996).

Biodegradation of aromatic compounds also commonly occurs in anoxic conditions and has been observed in proteobacteria, sulfate reducing bacteria, iron-reducing bacteria and fermentative bacteria. It requires an entirely different strategy in which all oxygen-dependent steps must be replaced by new reactions and alternative central intermediates (Heider and Fuchs, 1997). It typically occurs via reduction to break aromaticity followed by hydrolytic ring opening (Harwood et al., 1998). The first step in the anaerobic metabolism of aromatic compounds generally begins with the transformation of the compound to an acyl coenzyme A (acyl-coA) derivative (of which benzoyl-CoA is the most common), which could then be reduced (Schuhle et al., 2003). Anaerobic degradation of benzoate relies on the activation of benzoate by an ATP-dependent benzoate-coA ligase to produce benzoyl-coA (Valderrama et al., 2012). From benzoyl-coA there are two mechanisms to degrade the molecule anaerobically; one possible anaerobic pathway involves ring hydrolysis of cyclic 6-ketoxycyclohex-1-ene-1-carboxyl-coA while the other proceeds through ring hydrolysis of 2-ketocyclohexane-1-carboxyl-coA. The two anaerobic pathways then converge at 3-hydroxy-pimeloyl-coA and is then converted into glutaryl-CoA, which then undergoes glutaryl-coA degradation to produce two molecules of acetyl-coA. Overall, benzoyl-coA metabolism produces three molecules of acetyl-coA and one molecule of CO_2 (Harwood et al., 1998).

Benzoate degradation is particularly interesting to examine because many proteins and reactions have been added to the pathway in Mtb through the methods described thus far. Of these annotated reactions, PT had marked many of these as pathway holes, or reactions where Mtb either needed the enzyme catalysing that reaction or used the host metabolism to complete that reaction. Also, since Mtb can survive in both aerobic and anaerobic conditions, benzoate degradation is of potential interest for investigation of potential drug targets during the persistent (hypoxic) phase (Valderrama et al., 2012). Thus, the pathway was investigated for notable added reactions.

The KEGG diagram for the *M. tuberculosis* H37Rv benzoate degradation pathway shows only nine reactions as being annotated with enzymes within the genome. Of these, eight enzymes also catalyse reactions in other pathways leaving only one enzyme that is

unique to benzoate degradation. MetaCyc predicts the existence of multiple pathways within the KEGG pathway diagram for benzoate degradation. These include protocatechuate degradation II, benzoate degradation II (anaerobic), benzoyl-coA degradation III, catechol degradation II (which includes catechol degradation to 2-oxopent-4-enoate II and 2-oxopentenoate degradation), catechol degradation to β-ketoadipate, 3-oxoadipate degradation, benzoate degradation I (aerobic) and glutaryl-coA degradation. These MetaCyc pathways, among others, are all part of the benzoate degradation pathway in KEGG and have been used to break down the benzoate degradation pathway into smaller pieces and describe the portions of the pathway that have been found to have enzymes catalysing their reactions. The KEGG diagram for this pathway along with superimposed MetaCyc pathways is shown in Figure 4.2.

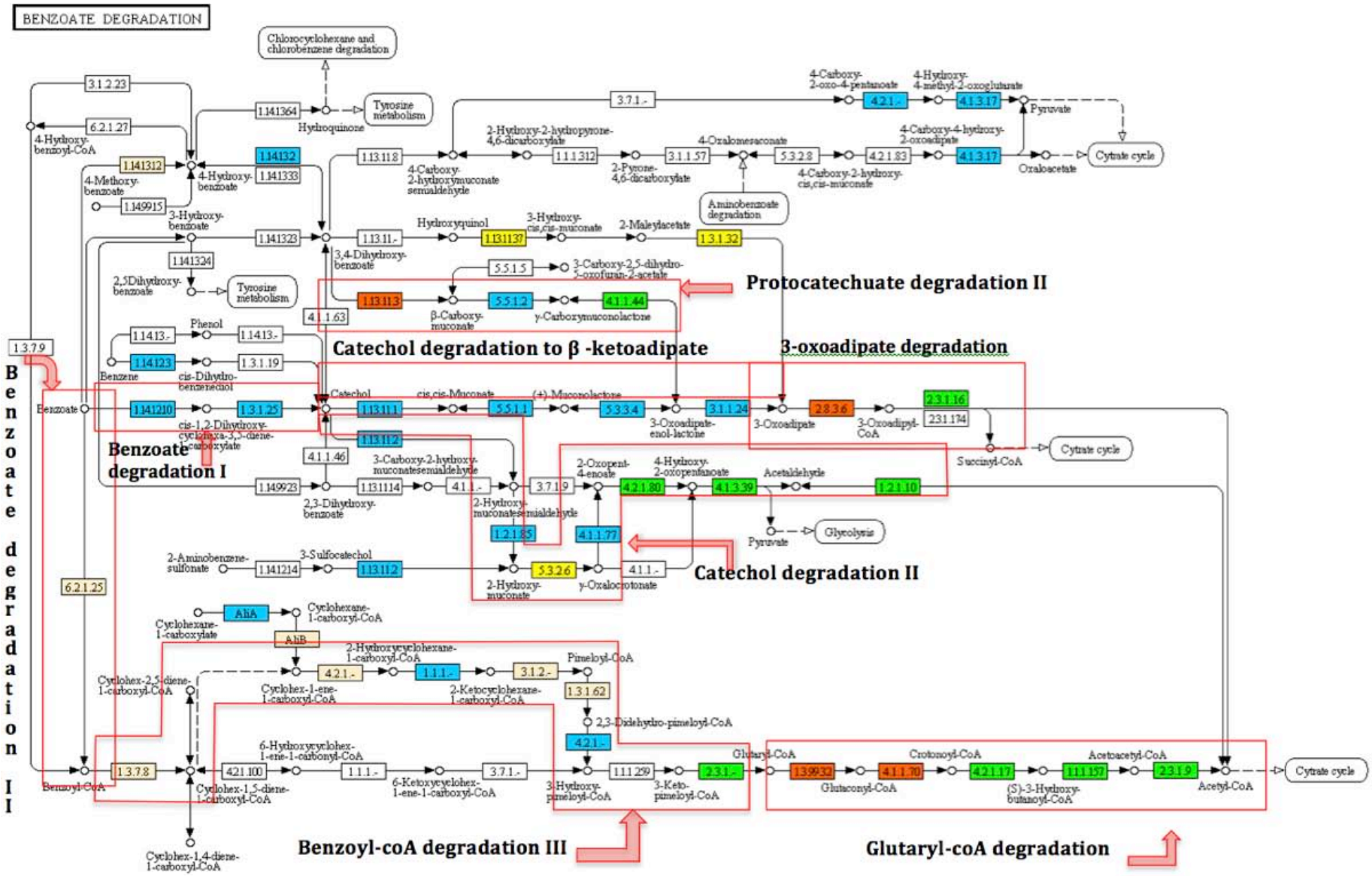


Figure 4.2: The KEGG pathway for *M. tuberculosis* H37Rv benzoate degradation. Red boxes show MetaCyc pathways. The boxes contain EC numbers and are different colours. Green indicates reactions already annotated in KEGG, blue are those annotated in this project, yellow are reactions with proteins in other organisms but for which no BLASTP matches could be found, orange are reactions classified as missing while pale orange shows possible additional missing reactions (Caspi et al., 2012; Kanehisa et al., 2004; Ogata et al., 1999).

As can be seen in the diagram, many additions were made to the pathway. The aerobic degradation of benzoate has been observed through three separate pathways: 'catechol degradation to β-ketoadipate' and 'protocatechuate degradation II' coupled with '3-oxoadipate degradation' and 'catechol degradation II.' Evidence for the existence of these pathways varies, but the 'benzoate degradation I' pathway is necessary for degradation through the intermediate catechol.

The 'benzoate degradation I' pathway has been completed using BLASTP matches with the toluate 1,2-dioxygenase electron transfer unit of *M. smegmatis* MC2 155 and the 1,6-dihydrocyclohexa-2,4-diene-1-carboxylate dehydrogenase of *R. jostii* RHA1. The e-values for these protein matches are $1e^{-47}$ and $5e^{-29}$, respectively, signalling relatively high confidence matches. This pathway is central for the aerobic degradation of benzoate through catechol and thus its annotation could be highly important for the functional characterisation of *M. tuberculosis* H37Rv.

The other aerobic intermediate, protocatechuate, or 3,4-hydroxybenzoate, is shown within the benzoate degradation pathway to be produced via a benzoate 4-hydroxylase from benzoate. However, no matches were found to annotate this reaction. Additionally, none of the organisms used in the BLASTP analyses showed proteins for this reaction. The production of protocatechuate is also possible from chorismate and the enzyme catalysing this reaction has been found in *M. tuberculosis* H37Rv, identified as a chorismate pyruvate-lyase (Stadthagen et al., 2005). Because the production of protocatechuate can occur through alternative means, it is possible that the reaction using the substrate benzoate to produce protocatechuate does not occur in *M. tuberculosis* H37Rv.

The convergent 'catechol degradation to β-ketoadipate' and 'protocatechuate degradation II' pathways were also partially annotated. Most of the reactions proceeding from catechol had previously been classified as pathway holes in Pathway Tools, and these holes have now been completed through matches with proteins in both *M. smegmatis* MC2 155 and *R. jostii* RHA1. All four enzymes have been found for the pathway proceeding from catechol, while two of three enzymes in the pathway from protocatechuate are annotated. A protocatechuate: oxygen 3,4-oxidoreductase has still yet to be identified for this pathway. Lastly, one additional enzyme, a β-

ketoadipate:succinyl CoA transferase, has also not yet been identified within the '3-oxoadipate degradation' pathway. Coupled with the 'benzoate degradation I' and '3-oxoadipate degradation' pathways, the 'catechol degradation to β-ketoadipate' pathway allows for the aerobic degradation of benzoate, and there is additional evidence that 'protocatechuate degradation II' also exists in *M. tuberculosis* H37Rv although it is unknown whether this pathway involves the degradation of benzoate itself as a source of energy.

The anaerobic degradation of benzoate is shown as only partially completed for Mtb in the KEGG diagram. Certain reactions have been annotated with enzymes, but many have not been identified. Of those that are currently annotated, all of them exist in other pathways, meaning that their existence in this pathway could simply be an artefact. Thus, it is still questionable as to whether Mtb can anaerobically degrade benzoate. The first major step in the anaerobic degradation is the transformation of benzoate to benzoyl-coA (Schuhle et al., 2003). No benzoate-coA ligase was found in Mtb or for any of the compared closely related organisms, and thus this might indicate the lack of a mechanism for anaerobic benzoate degradation in Mtb. The main anaerobic benzoate degradation pathway, labelled benzoyl-coA degradation III, shows only three identified enzymes of the eight required by the pathway. Among the missing enzymes include those required for ring reduction and cleavage. The last section in the anaerobic mechanism for benzoate degradation is glutaryl-coA degradation. This pathway is annotated with three of the five enzymes required, all of which function in other pathways. While the results shown demonstrate the existence of some enzymes catalysing reactions within the anaerobic degradation of benzoate, there is by no means conclusive evidence that this pathway exists in Mtb. In fact, due to the lack of enzymes found catalysing anaerobic benzoate degradation it is quite possible that Mtb cannot degrade benzoate in the absence of oxygen.

Overall the results show that Mtb is able to degrade benzoate aerobically. This is a common ability among soil bacteria, a group to which most species within the genus *Mycobacterium* typically belong (Harwood and Parales, 1996). These results suggest two possibilities as to the importance of benzoate degradation in Mtb: first, aromatic compound degradation could be an important metabolic ability within the host;

alternatively the ability to degrade benzoate could be an evolutionary artefact because of its evolution from a genus of which most species are soil bacteria.

In order to find additional evidence for the newly annotated proteins, Ensembl Bacteria was used to locate their encoding genes on the chromosome (Flicek et al., 2014). These proteins are shown in Figure 4.3.

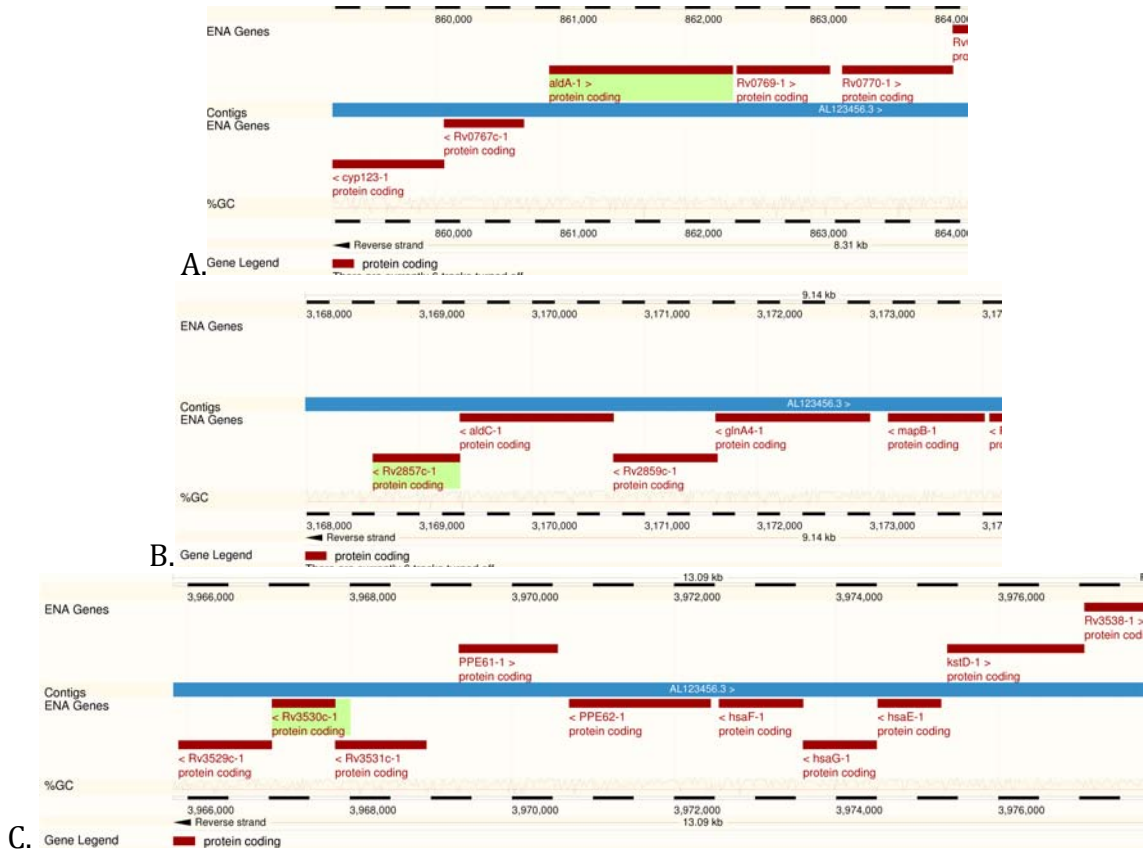


Figure 4.3 Figures showing the chromosome location of four newly annotated Mtb H37Rv genes to two reactions that have one reaction in between. Image A shows the locations of two genes encoding proteins (Rv0768 (P71823, EC:1.2.1.85) and Rv0769 (P71824, EC:1.3.1.25)) for the two reactions while image B shows the locations of genes encoding two different annotated proteins (Rv2857c (O33339, EC:1.3.1.25) and Rv2858c (O33340, EC:1.2.1.85)) for the same two reactions. Image C shows another of the newly annotated proteins Rv3530c (P71871) close to three already annotated in KEGG proteins Rv3534c (P71867, EC:4.1.3.39), Rv3535c (P71866, EC:1.2.1.10) and Rv3536c (P71865, EC:4.2.1.80) (Flicek et al., 2014).

These genes in Figure 4.3A and B encode multiple BLASTP matches for the reactions with the EC numbers 1.3.1.25 (in "benzoate degradation I') and 1.2.1.85 (in 'catechol degradation II') (see Figure 4.2). All four of these proteins were annotated using BLASTP matches with proteins in other organisms and appear to occur in pairs on the genome. In addition, EC:1.3.1.25 sits just four genes away from three consecutive previously characterized enzymes in this pathway (Figure 4.3C). The locations of these proteins suggest that they may function in operons, or are regulated by the same

means. These results provide evidence for the accuracy of the additional annotated proteins in Mtb H37Rv.

4.2.2 Phenylalanine Metabolism

Amino acid metabolism is an interesting category of pathways to examine, because most of the genes involved in amino acid biosynthesis are highly conserved across the genomes, with only *M. leprae* showing some loss of genes (Marri et al., 2006). Amino acids are protein building blocks and have vital functions in cell metabolism. The amino acid phenylalanine is a precursor of tyrosine and is essential for its synthesis (Wu, 2009).

Phenylalanine concentrations in mycobacteria have been found to be relatively low compared to other amino acids and compared to its reported concentrations in vertebrates and other microorganisms (Ginsburg et al., 1956). The phenylalanine metabolism pathway involves the production of various compounds including phenylpropanoids and benzenoids, most common in plants but shown to have counterparts in bacteria; these can have important roles in defence and signalling. It has been found that *Streptomyces* can even produce benzoyl-coA in a plant-like manner from phenylalanine (Moore et al., 2002). Phenylacetate, produced from phenylalanine, is a common aromatic intermediate that is often degraded into common metabolites in order to produce energy (Luengo et al., 2001).

The phenylalanine metabolism pathway in KEGG comprises a number of branches and parts that can be identified with MetaCyc identifiers. Those marked as potentially existing in *M. tuberculosis* H37Rv include 'phenylalanine degradation I (aerobic)', 'phenylethanol biosynthesis', 'phenylethylamine degradation I', 'phenylacetate degradation I (aerobic)', 'cinnamate and 3-hydroxycinnamate degradation to 2-oxopent-4-enoate', phenylpropanoid biosynthesis, initial reactions', 'phenylpropanoid biosynthesis' and 'capsaicin biosynthesis'.

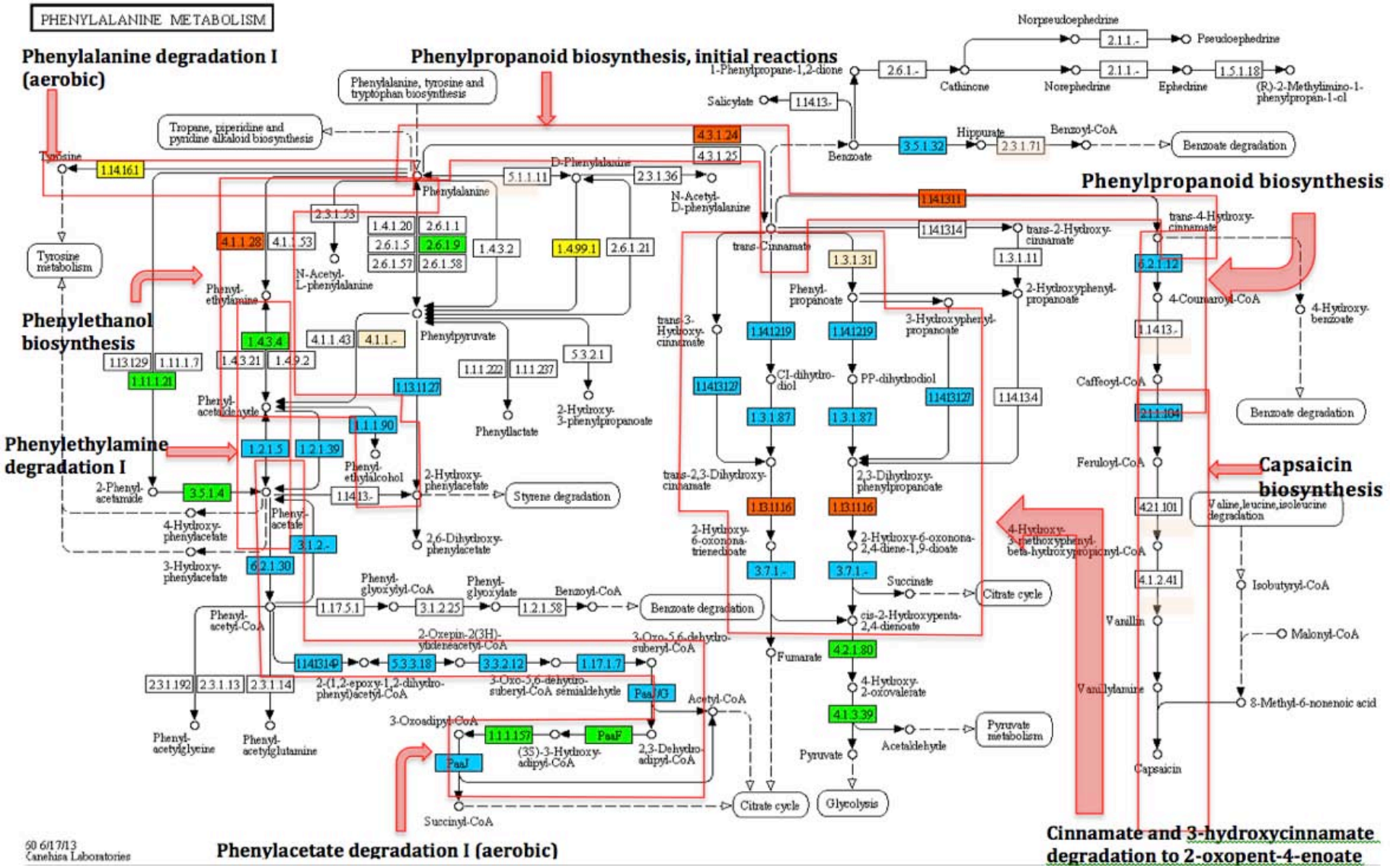


Figure 4.4: The KEGG pathway for M. tuberculosis H37Rv phenylalanine metabolism. Red boxes show MetaCyc pathways. The boxes contain EC numbers and are different colours. Green indicates reactions already annotated in KEGG, blue are those annotated in this project, yellow are reactions with proteins in other organisms found using BLASTP but for which no matches could be found in Mtb, orange are reactions classified as missing while pale orange shows possible additional missing reactions without additional evidence (for example, from Pathway Tools or existing in closely related organisms) that they are missing (Caspi et al., 2012; Kanehisa et al., 2004; Ogata et al., 1999).

Many Mtb proteins were identified for reactions in the phenylalanine metabolism pathway. The section labelled 'phenylethanol biosynthesis' originally showed an enzyme for one reaction, but here one more reaction was characterised using BLASTP, leaving one reaction unannotated. In *E. coli* K12 'phenylethanol biosynthesis' can be used as a sole source of energy by degrading 2-phenylethylamine to phenylacetic acid. The reaction converting phenylethylamine to phenylacetaldehyde also functions within the 'phenylethylamine degradation I' pathway, along with the subsequent conversion to phenylacetate (Parrott et al., 1987). Phenylethyl alcohol is mainly classified as an aromatic plant product and has been reported to exhibit antibacterial action, particularly in gram-negative bacteria (Lilley and Brewer, 1953). Since this enzyme functions in four other pathways as well, it is possible that its annotation within phenylalanine metabolism is simply an artefact. The missing enzyme shown within the pathway is a general aromatic amino acid decarboxylase which also degrades tyrosine, histidine and tryptophan (Kanehisa et al. 2012). Thus it could be an important enzyme that remains to be elucidated.

Phenylacetate is an important intermediate in the degradation of multiple aromatic compounds (Luengo et al., 2001). Thus phenylacetate degradation is an important part of the metabolism of aromatic compounds for use as energy. It can occur both aerobically and anaerobically, although no enzymes were found for the anaerobic degradation of phenylacetate, and so it probably does not occur in Mtb. Aerobic degradation of phenylacetate in Mtb has been elucidated; importantly, intermediates of this pathway can contribute to virulence, evidence of which has been found in multiple bacterial species and suggested in *M. abscessus*. Accumulation of the early products of phenylacetate degradation can have toxic effects on the host (Teufel et al., 2010). Phenylacetate degradation involves multiple CoA thioesters, epoxide formation, isomerisation to an oxepin and hydrolytic ring cleavage, resulting in succinyl-CoA and acetyl-CoA (Teufel et al., 2010). This is very similar to one of the pathways used by bacteria to aerobically (hybrid aerobic) degrade benzoate (Teufel et al., 2010), and could be important in Mtb. However, the entire aerobic hybrid phenylacetate degradation pathway could be characterised in Mtb using BLASTP matches primarily with *R. jostii* RHA 1 but also with *M. smegmatis* MC2 155 and *N. farcinica.* Of the ten enzymatic reactions necessary for aerobic degradation of phenylacetate only two

enzymes in Mtb were previously annotated. All of the eight remaining reactions were characterised using BLASTP matches, thus completing the pathway. These results show that Mtb might utilise phenylalanine as an energy source.

The 'phenylpropanoid biosynthesis, initial reactions' includes two reactions, neither of which could be characterised. Phenylpropanoids are compounds typically produced by flowering plants, and thus they are most likely not synthesised in Mtb (Caspi et al., 2012). However, the reactions following the initial reactions, including 'cinnamate and 3-hydroxycinnamate degradation to 2-oxopent-4-enoate' and 'phenylpropanoid biosynthesis' have been characterised using BLASTP matches.

Matches have been found for eight of eleven reactions within the 'cinnamate and 3-hydroxycinnamate degradation to 2-oxopent-4-enoate' pathway. This pathway involves the degradation of phenylpropanoids, compounds found in abundance in nature as the breakdown products of plant materials and soil proteins (Caspi et al., 2012). Bacterial degradation plays an important role in the carbon recycling of phenylpropanoids in the environment (Vicuña, 1988). It is unclear whether phenylpropanoids would exist in the macrophage; however, since the biosynthesis of these compounds is typically attributed to plants, it is unlikely they are common in the host. The discovery of matches for enzymes involved in this pathway might therefore be due to evolutionary relics, since many species in the genus *Mycobacterium* are soil dwellers. The reactions following this pathway already have characterised enzymes in KEGG. However, these two reactions are common intermediates in aromatic degradation and thus could also be involved in other pathways.

Lastly, the initial phenylpropanoid biosynthesis reactions are followed by reactions in the 'phenylpropanoid biosynthesis' and 'capsaicin biosynthesis' pathways. Within the first pathway, one of two reactions has been characterised in Mtb using BLASTP matches. Since lignins are typically plant compounds associated with cellulose (Boerjan et al., 2003) and the enzyme found also functions in another pathway, it is probable that this pathway does not actually exist in Mtb. The 'capsaicin biosynthesis' pathway follows the 'phenylpropanoid biosynthesis' pathway and one of three reactions in this pathway has been characterised with a BLASTP match. Capsaicin has only been found in

the fruits of plants in the *Capsicum* genus so it is unlikely that this pathway actually exists in Mtb (Sukrasno and Yeoman, 1993).

Thus, 22 reactions of the phenylalanine pathway have been assigned enzymes from the genome of *M. tuberculosis* H37Rv using BLASTP matches. These matches show complete or almost complete pathways that enable the degradation of phenylalanine through the bacterial aerobic hybrid pathway as well as the degradation of the phenylpropanoid trans-cinnamate, both as sources of energy. To provide further evidence for the annotated proteins, Ensemble Bacteria was used to locate the genes encoding these proteins on the chromosome (Flicek et al., 2014). Images showing five of these gene locations are in Figure 4.5.

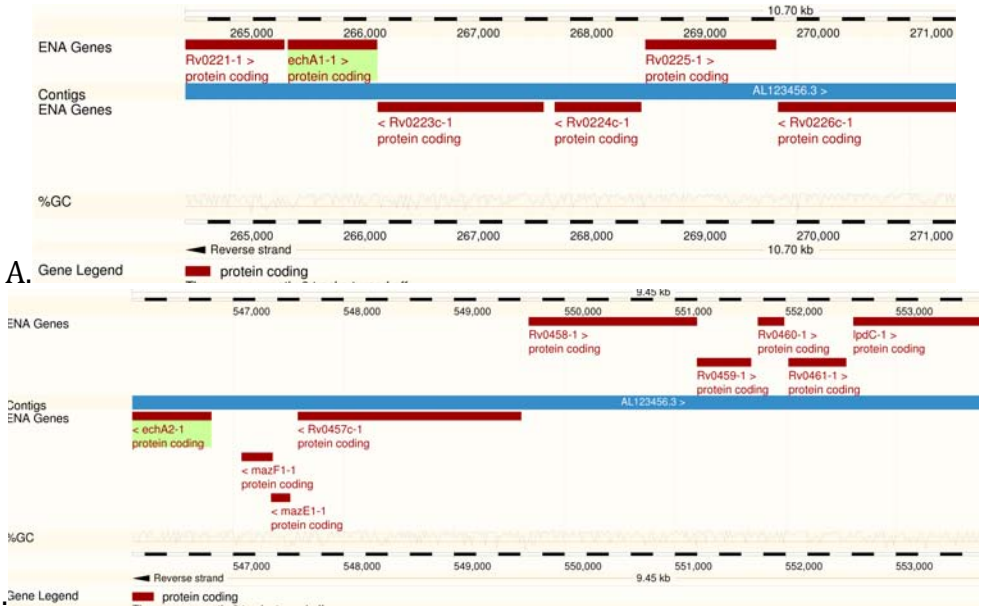


Figure 4.5 These images show the chromosome locations for genes encoding four proteins in Mtb H37Rv. Image A shows two genes Rv0222 (P96404, PaaF) and Rv0223c (P96405, EC:1.17.1.7) and image B shows two other genes (Rv0456c (O07179, PaaF) and Rv0458 (P63937, EC:1.17.1.7)). (Flicek et al., 2014).

Figure 4.5 shows genes encoding proteins in reactions catalysed by enzymes EC:1.17.1.7 and PaaF in 'phenylacetate degradation I (aerobic)', as shown in Figure 4.4. Two proteins, P96404 (Rv0222) and O07179 (Rv0456c), are already annotated in KEGG, while P96405 (Rv0223c) and P63937 (Rv0458) have been annotated with BLASTP matches. The close locations of these proteins suggest that are co-regulated and this provides evidence for their activity in the same pathway. The locations also lend support for the annotations derived from the BLASTP matches.

4.2.3 Glyoxylate and Dicarboxylate Metabolism

Central carbon metabolism, including glycolysis/gluconeogenesis, the pentose phosphate shunt and tricarboxylic acid (TCA) cycle, has been identified as a key determinant of the pathogenicity of Mtb (Rhee et al., 2011). While microbes can use non-carbohydrate sources of carbon and energy including fatty acids, lipids, amino acids, nucleotides and others, the degradation of these alternative sources typically proceeds through intermediates which are then processed through the central carbohydrate metabolism pathways (Moat et al., 2003). Because Mtb resides in such a narrow environmental niche within the human host, carbon metabolism is particularly important for its survival and growth; the bacteria can simultaneously process multiple substrates of carbon, an ability not seen in many other bacterial species (de Carvalho et al., 2010).

Glyoxylate and dicarboxylate metabolism allows microorganisms to grow on acetate as a sole source of carbon during growth (Moat et al., 2003). Additionally, this pathway is part of those of central metabolism, which together produce the 13 precursor metabolites for all cellular biosynthesis. This pathway can also generate energy for growth (Caspi et al., 2012).

MetaCyc pathways which form sections of the KEGG 'glyoxylate and dicarboxylate metabolism' pathway include the 'glyoxylate cycle,' 'glycolate and glyoxylate degradation I,' 'glycolate and glyoxylate degradation II,' 'oxalate degradation III,' 'oxalate degradation V,' 'formaldehyde assimilation I (serine pathway),' 'ethylmalonyl pathway' and 'methylmalonyl pathway'. These pathways are shown in Figure 4.6.

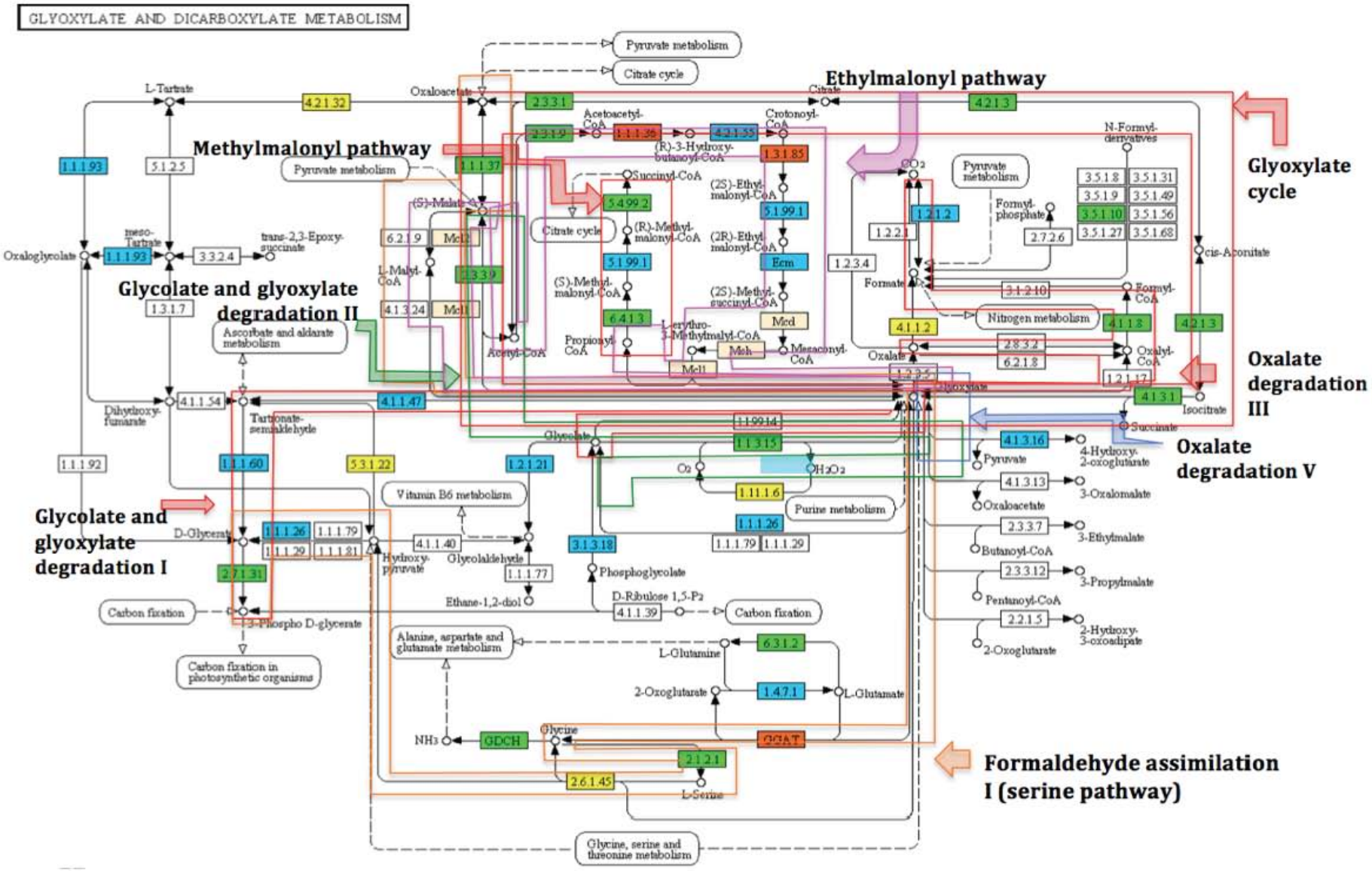


Figure 4.6: The KEGG pathway for M. tuberculosis H37Rv glyoxylate and dicarboxylate metabolism. Red boxes show MetaCyc pathways. The boxes contain EC numbers and are different colours. Green indicates reactions already annotated in KEGG, blue are those annotated in this project, yellow are reactions with proteins in other organisms but for which no BLASTP matches could be found in Mtb, orange are reactions classified as missing while pale orange shows possible additional missing reactions without further evidence (such as pathway holes in PT or characterised in closely related organisms) (Caspi et al., 2012; Kanehisa et al., 2004; Ogata et al., 1999).

One of the most important metabolic functions within this pathway involves the reactions belonging to the 'glyoxylate cycle', part of the central carbon metabolism of microorganisms. This cycle utilises acetate for growth and energy, and results in the formation of malate from two molecules of acetate. Both glucose and fatty acids can be converted into acetate for use in the glyoxylate cycle, which is particularly important for use of fatty acids as substrates for growth. The glyoxylate cycle resembles the tricarboxylic acid (TCA) cycle except that it bypasses those steps of the cycle which lead to a loss of carbon in the form of CO_2 (Moat et al., 2003). The cycle also includes the enzyme isocitrate lyase, shown to be essential for *in vivo* growth and virulence of Mtb (Munoz-Elias and McKinney, 2005). All of these reactions have already been characterised with enzymes in KEGG, and thus no additions were made. However, MetaCyc shows two different enzymes for malate synthase (EC:2.3.3.9), one of which, malate synthase A, functions in the 'glyoxylate cycle' in *E. coli* and the other, malate synthase G, functions in the 'glycolate and glyoxylate degradation II' pathway, discussed later (Clark and Cronan, 2005). Only the gene encoding malate synthase G has been identified in Mtb, meaning that this step is actually missing or uncharacterised for the glyoxylate cycle. No protein for malate synthase A could be located within the genome of Mtb.

Even though alternative degradation pathways exist for glycolate and glyoxylate, the 'glycolate and glyoxylate degradation I' pathway is essential for growth on either compound in *E. coli* possibly because its products cycle into glycolysis. Through this pathway *E. coli* can use either of these two substrates as the lone source of carbon and energy. The glyoxylate utilised in this degradation pathway can come from either exogenous or endogenous (i.e.- via other pathways) sources (Clark and Cronan, 2005). Of the four reactions belonging to this pathway in Mtb, two have been characterised using the BLASTP matches, one is already characterised in KEGG and the fourth has not been identified for any organism in KEGG. Of the two BLASTP matches, the enzyme found matching glyoxylate carboligase (EC:4.1.1.47) exhibits a low e-value [$3e^{-86}$] but is described as an acetolactate synthase; therefore the match could be inaccurate for this reaction. The fourth reaction, for which an enzyme does not exist in KEGG for any organism, could be characterised based on data from MetaCyc and UniProt. Using the protein name found in MetaCyc, a UniProt search found one of three subunits of the

enzyme required to catalyse this reaction. Thus, this reaction is in fact partially characterised for Mtb. Although a protein catalysing the reaction between glyoxylate and tartronate semialdehyde has yet to be found with confidence, the characterisation of the other enzymes within this pathway would suggest that the pathway does exist in Mtb. This enables the bacteria to utilise glyoxylate as a sole source of energy and carbon when nutrients are scarce, such as probably occurs within the macrophage.

The 'glycolate and glyoxylate degradation II' pathway also functions in degrading glyoxylate as a source of carbon and energy; however, it is not essential in *E. coli* and instead cycles into the TCA cycle rather than glycolysis (Clark and Cronan, 2005). This pathway shares the first step with 'glycolate and glyoxylate degradation I', which was characterised using the protein name and function in order to identify a gene encoding one of three subunits of the enzyme catalysing this reaction. The second reaction in this pathway was already characterised in KEGG. Thus, this pathway is only partially complete, requiring the identification of the genes encoding the remaining two subunits of glycolate oxidase in order to be completed. None of the closely related organisms had characterised genes for these two subunits and thus no attempts at BLASTP matches could be performed.

Most of the 'formaldehyde assimilation I (serine pathway)' pathway is included in the KEGG 'glyoxylate and dicarboxylate metabolism' diagram apart from two steps. Formaldehyde is a compound derived from methanol via oxidation. In one study, all mycobacteria tested were found to be methylotrophic, meaning they could grow solely on single-carbon sources of carbon and energy such as methanol and carbon monoxide; this is unique among taxonomic groups of bacteria. The singular outlier of the study, Mtb was only able to grow on carbon monoxide (CO) and not on methanol (Park et al., 2003). Of the nine reactions from this pathway shown in the KEGG diagram, four were previously characterised in KEGG, one was characterised using BLASTP matches and four remain uncharacterised. Of these four uncharacterised reactions, two are catalysed by the same enzyme in a coupled reaction. Thus only about half the pathway has been characterised, and it is unclear whether this pathway is functional in Mtb. Additionally, 'formaldehyde assimilation I (serine pathway)' begins by utilising formaldehyde as a substrate, itself produced from methane using either a methanol dehydrogenase or

ribulose bisphosphate carboxylase/oxygenase (RuBisCO). A pathway enabling the production of formaldehyde from CO was not found during a literature review; the use of CO as a source of carbon and energy rather proceeds via a carbon monoxide dehydrogenase to produce CO_2, which then progresses into the TCA cycle. The inability of Mtb to grow on methanol combined with the four uncharacterised reactions suggests that this pathway might not be functional in Mtb. Because many other species within *Mycobacterium* are able to grow on methanol, perhaps the existence of enzymes catalysing the characterised reactions is simply an evolutionary artefact rather than a sign that the pathway is functional.

The 'ethylmalonyl pathway' is an alternative pathway for the metabolism of fatty acids via acetate. As opposed to the 'glyoxylate cycle', the 'ethylmalonyl pathway' enables organisms to metabolise acetate when isocitrate lyase (ICL), a key enzyme in the 'glyoxylate cycle', is not available. This pathway functions in organisms that are unable to produce ICL and proceeds through acetoacetyl-CoA and other intermediates to produce glyoxylate, recycled back into (S)-malate, and propionyl-CoA, which is carboxylated to succinate via the 'methylmalonyl pathway' (Alber et al., 2006). The key step in this pathway is the conversion from crotonoyl-CoA to (2S)-ethylmalonyl-CoA, catalysed by crotonyl-CoA carboxylase/reductase (Erb et al., 2007). Of the 13 reactions within the ethylmalonyl and methylmalonyl pathways, four were already characterised in Mtb, four were characterised using BLASTP matches, and five are still uncharacterised. The key enzyme crotonoyl-CoA carboxylase/reductase was not found in Mtb or any of the examined closely related organisms. Since this pathway is only partially characterised, and since Mtb can utilise the glyoxylate cycle to metabolise two-carbon substrates, it is unclear whether this pathway is functional. While the 'ethylmalonyl pathway' is still incomplete, the 'methylmalonyl pathway' has been completed. This pathway succeeds the 'ethylmalonate pathway' and begins with propionyl-CoA. Mtb can obtain propionate from its environment and uses it to form methylmalonate, an important component in the synthesis of mycobacterial cell wall lipids (Matsunaga et al., 2004). Thus, the completion of the 'methylmalonyl pathway' is not necessarily evidence of the existence of a functional 'ethylmalonyl pathway', and it is unclear whether the latter operates in Mtb.

The ‘oxalate degradation V’ and ‘oxalate degradation III’ pathways enable the metabolism of oxalate, a compound commonly found in the soil and produced by some plants and fungi (Allison et al., 1995). Two reactions producing oxalate from glyoxylate are shown in KEGG but neither of them have an assigned enzyme in any organism in the KEGG database. Thus, it is unknown whether Mtb might produce oxalate from glyoxylate, metabolise exogenous sources or not use oxalate at all. Two different pathways show possible evidence that oxalate degradation exists in Mtb for oxalate degradation, with ‘oxalate degradation V’ using an oxalate decarboxylase to produce formate directly from oxalate and ‘oxalate degradation III’ utilising an oxalate-CoA transferase and oxalyl-CoA decarboxylase to convert oxalate first to oxalyl-CoA, then formyl-CoA and finally to formate. However, both of these pathways remain largely uncharacterised in KEGG, even for other organisms; thus BLASTP could not be used to find matches for these enzymes in the genome of Mtb. Since no reactions could be characterised for ‘oxalate degradation III’ and only one reaction is already characterised in KEGG, it is unclear whether this pathway is functional. While several other organisms contained the oxalate decarboxylase catalysing the reaction of ‘oxalate degradation V’, no matches could be identified in Mtb. Additionally, because oxalate is a common compound in soil and many *Mycobacteria* are soil-dwellers, it is possible that the existence of portions of these pathways is an evolutionary artefact.

Overall, several reactions within the ‘glyoxylate and dicarboxylate metabolism’ pathway were characterised in Mtb using BLASTP matches. These matches reveal evidence for the existence of several MetaCyc pathways including ‘glycolate and glyoxylate degradation’ I and II, ‘formaldehyde assimilate I (serine pathway)’ and ‘methylmalonyl pathway’. Additionally there is partial but not strong evidence for the existence of the ‘methylmalonyl pathway’ and ‘oxalate degradation’ III and V. As before, Ensembl Bacteria was then used to find additional evidence for the newly annotated proteins (Flicek et al., 2014). Images shown in Figure 4.7 represent the locations of two newly annotated Mtb H37Rv proteins.

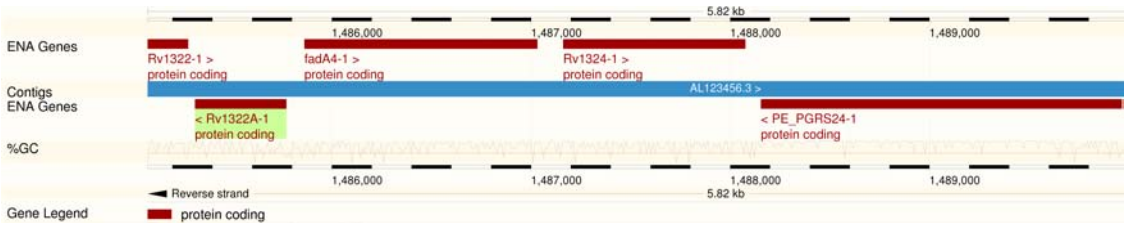


Figure 4.7 Image showing the location of two genes encoding proteins in the 'glyoxylate and dicarboxylate metabolism' pathway. It shows the genes Rv1322A (Q8VK36, EC:5.1.99.1) and Rv1323 (P66926, EC:2.3.1.9) (Flicek et al., 2014).

The proteins shown in Figure 4.7 function in the 'ethylmalonyl pathway' and 'methylmalonyl pathway'. Located next to one another on the chromosome, their expression is likely controlled by the same regulatory proteins. These are both newly annotated proteins, and although not co-located with others in the pathway they are co-located with each other.

4.3 Comparing Pathways between Organisms

The phylogenetic profile allowed for the comparison of protein numbers belonging to the different metabolic pathways over a large range of different organisms. By comparing Mtb with *M. leprae*, a pathogenic mycobacterial species with a highly reduced genome, and *E. coli,* often used as a model organism and thus well characterised functionally, certain interesting pathways have emerged. These interesting pathways include one or more of five particular characteristics: those in which many proteins have been added to both Mtb and *M. leprae,* those in which many proteins have been characterised for Mtb and not for *M. leprae,* those for which the pathway totals are very different between Mtb H37Rv and *E. coli,* those for which Mtb and *C. glutamicum,* a facultative anaerobe, share many homologs, and those in which Mtb and *H. sapiens* do not share many homologs.

Proportions were calculated for the number of homologous proteins for each organism in each of the metabolic pathways. These were calculated by taking the number of homologous proteins for each organism and dividing by the total number of proteins for that pathway in *M. tuberculosis* H37Rv. This was done so that all pathways could be compared on the same scale (for example, zero to one) for visualisation in the heat map. The pathways were then compared to one another by averaging the proportions of all organisms for each pathway. Those with the highest average are those pathways which are most conserved among the analysed genomes and vice versa for those with the

lowest averages. This should confirm or refute research examining the most and least conserved pathways among organisms. One study found that all examined genomes showed similar (conserved) pathways for amino acid biosynthesis, cofactor biosynthesis, nucleotide metabolism and macromolecule metabolism (Marri et al., 2006). The top five pathways according to average frequency include 'D-alanine metabolism' (metabolism of other amino acids), 'thiamine metabolism' (metabolism of cofactors and vitamins), 'D-glutamine and D-glutamate metabolism' (metabolism of other amino acids), 'lysine biosynthesis' (amino acid metabolism) and 'L-cysteine biosynthesis' (amino acid biosynthesis). Of these pathways, four belong to amino acid biosynthesis and metabolism; this aligns with the previous research (Marri et al., 2006). The fifth pathway, 'thiamine metabolism', belongs to the 'metabolism of cofactors and vitamins' category; genes involved in folic acid, pantothenate, pyridoxine and thiamine biosynthesis have also been shown to be highly conserved among *Mycobacterium* (Marri et al., 2006). These results support the methods used in so far as they agree with previous research regarding evolutionary conservation of pathways.

On the other hand, the pathways with the lowest averages show the least conserved pathways across the phylogenetic profile. The five pathways with the lowest average frequencies across the phylogenetic profile include 'glycerolipid metabolism' (lipid metabolism), 'ether lipid metabolism' (lipid metabolism), 'linoleic acid metabolism' (lipid metabolism), 'D-arginine and D-ornithine metabolism' (metabolism of other amino acids) and 'carotenoid biosynthesis' (metabolism of terpenoids and polyketides). Lipid metabolism, cell wall proteins and polyketide synthases showed wide variation across strains of *Mycobacterium*, and both lipid metabolism and cell wall proteins could be related to virulence as many lipids function in the cell membrane, the interface between host and pathogen (Marri et al., 2006). Many of the low average frequency pathways found here belong to lipid metabolism, showing that these pathways differ quite extensively not just among mycobacteria but also across the entire phylogenetic profile. Three of these bottom five have very few characterised proteins (between one and five in Mtb out of two to twenty in the KEGG reference pathway) and this could have skewed the results. However, when looking at the bottom 16 pathways in terms of average frequencies, six of those pathways belong to the lipid metabolism category. Of the remaining, three are part of the metabolism of terpenoids and polyketides, two are

part of xenobiotics biodegradation and metabolism, and one pathway each belongs to carbohydrate metabolism, metabolism of cofactors and vitamins, and metabolism of other amino acids. The other two pathways are UniPathway pathways added into the profile in order to increase coverage and include cell wall polysaccharide biosynthesis and mycolic acid biosynthesis. Pathways with low conservation across the phylogenetic profile are particularly important to examine because they have potential as future drug targets. For example, drugs could possibly target an Mtb pathway that does not have protein homology with pathways in *H. sapiens* without causing disruption of human metabolism. These pathways with low average frequencies were then selected for further examination.

In addition to the bar graphs created to compare numbers of proteins characterised for Mtb*, M. leprae, C. glutamicum, E. coli,* and *H. sapiens* (Figure 3.3), which includes the protein counts for both KEGG pathways as well as those in the profile, a heat map was created of the phylogenetic profile as shown in Figure 3.1 (some organisms were removed to fit the heat map onto two pages). This heat map shows the number of proteins belonging to each pathway for each organism as a proportion of the number of proteins for each pathway in *M. tuberculosis* H37Rv. Thus each of these proportions are calculated out of a score of one and shown across the list of 392 organisms. The two figures enable a visual interpretation of the data to find pathways that might be possible drug targets.

Pathways matching each of the five characteristics were found using the heat map (Figure 3.1), frequency comparison graph (Figure 3.2) and bar graph (Figure 3.3). Many of these pathways satisfied more than one of the desired characteristics for the purpose of this study.

4.3.1 Many Proteins Added to Both *M. tuberculosis* and *M. leprae*

Two pathways for which a number of proteins were added in both Mtb H37Rv and *M. leprae* TN are 'arginine and proline metabolism' and 'glycine, serine and threonine metabolism'. Both these pathways show relatively equal numbers of proteins added in the two species of mycobacteria. These pathways are part of 'amino acid metabolism', and most of the genes involved in amino acid biosynthesis are highly conserved across the genomes (Marri et al., 2006). The results shown in Figure 3.2 for these two

pathways suggest that the characterised proteins found might be functionally important for mycobacteria. Additionally, the results from the compiled phylogenetic profile shown in Figure 3.1 show relatively high levels of homology for proteins involved in 'arginine and proline metabolism'. The frequency of homology is lower for 'glycine, serine and threonine metabolism' but still shows relatively high homology. Arginine degradation is a common ability among organisms from all kingdoms, and as such performs an important role in organisms (Jenkinson et al., 1996). Arginine and proline can both be degraded via glutamate or alternative pathways as a source of carbon or nitrogen. Previous studies have shown the existence of arginases in mycobacteria and the discovery of additional proteins in the pathway are likely the result of the characterisation of additional segments of the pathway (Zeller et al., 1954). These additions do not show homology with either *E. coli* or *H. sapiens* and thus could be possible drug targets. 'Glycine, serine and threonine metabolism', as another amino acid pathway, also has important functions and can be used as a source of carbon. Serine is produced from an intermediate of glycolysis, which then is transformed into glycine; threonine is produced from aspartate. Glycine can also be produced from betaine and can be used as a sole source of carbon and nitrogen (Smith et al., 1988). Threonine itself is indispensable, can proceed through both aerobic and anaerobic pathways and can be used as a solitary substrate for growth (Sawers, 1998). Therefore, all of these amino acid metabolism pathways can be used as sources for growth, and since they share many proteins with *M. leprae*, might be essential for growth and survival. This essentiality could make these pathways potential new drug targets.

Many proteins have also been added to the 'glycerolipid metabolism' pathway in both Mtb and *M. leprae.* Many bacteria can directly absorb glycerol from the environment to use as a source of carbon and energy. While some bacteria have been found to grow on glycerol aerobically, others can metabolise the compound anaerobically (Rush et al., 1957). Additionally, phospholipids are important components of the membrane and cell wall of Mtb; metabolism of glycerol proceeds through phosphoglycerides and can be important in building constituents of cell wall proteins. Proteins that are part of this pathway could be important drug targets for multiple reasons. First, Mtb shifts from a carbohydrate-based to a lipid-based metabolism while within the macrophage (Yang et al., 2011). Second, Mtb shares many orthologous proteins with *M. leprae*, suggesting

that these proteins would be essential for growth in the host. Lastly, glycerolipid metabolism has a very low number of orthologous proteins across the phylogenetic profile meaning it likely does not share many enzymes with other organisms and if drugs were to target these enzymes then they would not simultaneously affect the metabolism of human cells or other organisms within the host. All of these factors warrant a closer examination of this pathway and the newly characterised proteins for potential targets of new drugs.

4.3.2 Many Proteins Added to *M. tuberculosis* But Not *M. leprae*

'Fructose and mannose metabolism' is a pathway in which the number of proteins characterised in Mtb is far higher than the number of proteins characterised in *M. leprae*. This pathway also shows very few homologous proteins between Mtb and *E. coli*. Additionally, when looking across the phylogenetic spectrum (Figure 3.1), this pathway shows relatively few homologous proteins except for very closely related organisms such as *Rhodococcus*. That there were many proteins added in Mtb that did not have homologs in the highly reduced genome of *M. leprae* suggests that these added proteins are not absolutely necessary for survival. In all mycobacteria mannose-containing glycolipids are an important component of cell walls, but have also been shown to be essential for growth and cell division in *M. smegmatis*. The same study also suggested that enzymes involved in the production and metabolism of mannose could be possible drug targets, especially due to the fact that they share few homologous genes with animal cells (Patterson et al., 2003). Figures 3.1 and 3.2 also show the low levels of homology between Mtb genes involved in fructose and mannose metabolism and genes from other organisms including *H. sapiens*. These results suggest that 'fructose and mannose metabolism' could be a potential drug target for preventing the growth of Mtb. On the other hand, the lack of orthologous genes in *M. leprae* could mean that these reactions are not necessary for survival and thus may not make good drug targets.

Another pathway in which many proteins were added to Mtb H37Rv but not to *M. leprae* TN is 'limonene and pinene degradation'. 'Limonene and pinene degradation' is part of the metabolism of terpenoids and polyketides category in KEGG; polyketides have been shown to vary greatly among the species within the MTBC and thus may have proteins unique to Mtb (Marri et al., 2006). This pathway is also important to note in

that it does not have many orthologs in either *E. coli* or *H. sapiens.* Limonene is an incredibly widespread compound produced by many different species of plants. *R. erythropolis*, a closely-related organism to Mtb, can grow on limonene as a sole source of carbon and energy (van der Werf et al., 1999). Pinene is also a common monoterpenoid produced by certain plants (Caspi et al., 2012). The phylogenetic profile shows that this pathway has the characteristics that are necessary for drug targets in terms of the lack of homology with *H. sapiens* and other organisms; however, the compounds upon which this pathway operates are not likely to be found within the macrophage. Thus, the pathway is probably not essential for Mtb H37Rv and may simply be an artefact due to the soil-dwelling nature of many mycobacteria.

4.3.3 Large Differences Between *M. tuberculosis* and *E. coli*

Pathways for which proteins differ significantly between Mtb and *E. coli* could also signify potential drug targets. If there are few orthologous proteins between these two organisms then there is a good chance that there are also many differences between Mtb and other organisms, including those likely to be affected by drugs given to a patient. Pathways that show few orthologous proteins between these two organisms include 'histidine metabolism', 'butanoate metabolism' and 'phenylalanine degradation'.

'Histidine metabolism' is another amino acid pathway and thus expected to be conserved across the phylogenetic profile. However, this does not seem to be the case. In fact, Figures 3.1 and 3.2 show very few orthologous proteins for this pathway. Part of this pathway involves the biosynthesis of histidine from PRPP, and is the same pathway for all organisms studied so far, with only small differences in some of the enzymes used (Caspi et al., 2012). This section of the pathway was already characterised in KEGG and thus no proteins were added. On the other hand, the metabolism of histidine can proceed via multiple pathways using different enzymes and histidine can be used as a source of nitrogen and carbon (Borek and Waelsch, 1953). If histidine metabolism is essential for growth then reactions of this histidine degradation pathway could be possible new drug targets. In fact, a recent study demonstrated the essentiality of histidine biosynthesis for growth and suggested that enzymes within this pathway would make attractive drug targets (Lunardi et al., 2013). The phylogenetic profile

strengthens the attractiveness of this pathway as a potential drug target, and further research could be done to determine whether these enzymes would be viable targets.

Another pathway in which *E. coli* and Mtb H37Rv share few orthologous proteins is 'butanoate metabolism'. Of the many proteins added to this pathway, parts of multiple pathways were characterised. These include aerobic and anaerobic pathways involved in carbon metabolism as a source of carbon and energy. Additional reactions have been characterised, allowing the partial completion of the interconversion of pyruvate and succinate as well as the production of butanoyl-CoA from acetoacetate through both acetoacetyl-CoA and (R)-3-hydroxybutanoate. The bar graph shows many added proteins for this pathway, with very few added proteins in *E. coli* and *H. sapiens.* The pathway also has a substantial number of added proteins for *M. leprae*. Even though some of the reactions belong to anaerobic pathways, *C. glutamicum* and Mtb share very few orthologous proteins. Additionally, Figure 3.2 shows 'butanoate metabolism' in the lower half of pathways, meaning that it is not very conserved evolutionarily. The results from the phylogenetic profile thus point to 'butanoate metabolism' as a pathway that merits further investigation as a potential drug target.

Another pathway that shows relatively few shared orthologs between Mtb and *E. coli* is 'phenylalanine metabolism'. As stated earlier, 'phenylalanine metabolism' can function as a sole source of carbon and energy (Caspi et al., 2012). Additionally, amino acid metabolism pathways are typically highly conserved among organisms (Marri et al., 2006). However, when observing this pathway across the phylogenetic profile, very few orthologs are found between Mtb and *E. coli.* Additionally, Figure 3.3 shows many orthologs between Mtb and *M. leprae* and few orthologs between Mtb and *H. sapiens.* These findings suggest first that these newly characterised proteins may be important aspects of Mtb metabolism. Secondly, the lack of orthologs between Mtb and *E. coli* and *H. sapiens* indicates that these enzymes might not be common either in the human body or in beneficial bacteria that reside within the host. Both of these factors warrant a closer look at the reactions of 'phenylalanine metabolism' as targets for potential new drugs.

Thus a number of pathways including 'histidine metabolism', 'butanoate metabolism' and 'phenylalanine metabolism' show relatively few orthologs between Mtb and *E. coli.*

The lack of orthologs between these two organisms provides pathways and reactions, which, if targeted by drugs, would not affect pathways and reactions essential for beneficial bacteria within the human body.

4.3.4 Many Similarities Between *M. tuberculosis* and *C. glutamicum*

C. glutamicum is a closely related bacteria to Mtb which functions as a facultative anaerobic bacterium (Nishimura et al., 2007). Since Mtb also survives in an anaerobic state during persistence in the host, comparison of the phylogenetic profile between this organism and Mtb could show pathways that are particularly important for bacterial survival in an anaerobic environment. Pathways that were observed to fulfil this characterisation are 'xylene degradation' and 'porphyrin and chlorophyll metabolism'.

Xylene is a polycyclic aromatic hydrocarbon that can be found in the environment, and is a major petrochemical. Another species of mycobacteria, *Mycobacterium austroafricanum* IFP 2012, can grow solely on xylene (Francois et al., 2002); *M. cosmeticum* has also been found to degrade benzene, toluene, ethylbenzene and xylene (Zhang et al., 2013). Xylene is degraded via reactions shared with the degradation of other polycyclic aromatic hydrocarbons such as toluene, benzene and ethylbenzene. Xylene degradation does not show a large number of orthologs, as seen in Figure 3.2; however, it is one of the few pathways which shows a greater number of orthologs than are characterised for that pathway in KEGG. The phylogenetic profile shows relatively low conservation of proteins across all organisms, as shown in Figure 3.1. The ability to degrade xylene has not been directly shown in Mtb, but these results suggest that this is functionally possible for the microorganism. However, its importance as an essential pathway for anaerobic metabolism is inconclusive in this case.

Another pathway in which Mtb and *C. glutamicum* share many orthologs is 'porphyrin and chlorophyll metabolism'. The pathway results in the production of cobalamin (vitamin B-12), one of the most structurally complex molecules found in nature (Martens et al., 2002). Mtb is one of a select group of bacteria that can synthesise cobalamin *de novo*, but it can also uptake cobalamin from the host environment. It has even been suggested that the organism can regulate core metabolic functions based on the availability of cobalamin (Gopinath et al., 2013). This similarly complex pathway

proceeds via both aerobic (late cobalt incorporation) and anaerobic pathways (early cobalt incorporation) (Warren et al., 2002), both of which have been added to through manual annotation and BLASTP matches. The aerobic pathway has been completed while the anaerobic pathway now has only two missing enzymes. In Figure 3.3 this pathway shows more orthologs for *C. glutamicum* than for *M. leprae*. On the other hand, Figure 3.1 shows that this pathway is relatively conserved across the phylogenetic profile. Because the number of orthologs between Mtb and *C. glutamicum* are quite high, this suggests that certain of these proteins are important in anaerobic metabolism. It has been proposed that there is a strong role for cobalamin in pathogenesis (Gopinath et al., 2013). Further examination of this pathway, in particular the anaerobic section of this pathway, could thus be warranted during the search for potential drug targets that would inhibit Mtb functioning in the persistent phase, during which the bacteria is able to survive in an anaerobic environment.

Both 'xylene degradation', including other pathways that are part of xenobiotics biodegradation and metabolism, and 'porphyrin and chlorophyll metabolism' show many shared proteins between Mtb and *C. glutamicum.* Due to this similarity, these pathways might possibly be targets for new drugs that attack Mtb during the persistent phase of infection.

4.3.5 Large Differences Between *M. tuberculosis* and *H. sapiens*

For reactions and pathways to be labelled as potential drug targets, they must not be catalysed by the same enzymes in Mtb and *H. sapiens.* Using the phylogenetic profile of Figure 3.1 and 3.2, pathways were noted which showed few orthologous proteins between these two organisms. Certain pathways were noted in this regard including 'biotin metabolism', 'glyoxylate and dicarboxylate metabolism', 'sulfur metabolism', 'geraniol degradation' and 'polycyclic aromatic hydrocarbon degradation'.

The 'biotin metabolism' pathway has only two reactions added; however, these reactions partially complete the biosynthesis of biotin, an essential cofactor for carboxyl group transfer enzymes required by all forms of life (Caspi et al., 2012). In *E. coli* biotin is synthesised through a number of reactions belonging to a slightly altered fatty acid synthetic pathway. These reactions add a methyl ester to the intermediates to disguise them so that they can be substrates for the fatty acid synthesis reactions. After two

reiterations of these fatty acid elongation cycles, pimeloyl-[acp] is produced, which then proceeds via 8-amino-7-oxononanoate to produce biotin in reactions that have already been characterised in KEGG for Mtb (Lin et al., 2010). In the phylogenetic profile this pathway was found to share few orthologs with *H. sapiens*, although it shares a large proportion with both *M. leprae* and *E. coli*. The high proportion of shared orthologs with *M. leprae* suggests that this pathway is important for growth and survival, a finding supported by the literature (Caspi et al., 2012; Lin et al., 2010). The low proportion of shared orthologs with *H. sapiens* suggests these organisms share few similarities with regards to the biosynthesis of biotin. These results indicate that reactions of this pathway might be potential drug targets for future research.

Another pathway that deserves further investigation is 'glyoxylate and dicarboxylate metabolism'. As mentioned previously this pathway is incredibly important for the metabolism of fatty acids as a source of carbon and energy (Moat et al., 2003). Additionally, these reactions are important for the use of glycolate and glyoxylate as sources of carbon and energy (Clark and Cronan, 2005). Figure 3.3 shows that only about one quarter of the Mtb proteins involved in this pathway have orthologs in *H. sapiens.* Those proteins shared between the two organisms mostly comprise the 'methylmalonyl pathway', 'glyoxylate cycle' and 'glycolate and glyoxylate degradation I' pathway. Of these three pathways, only the 'methylmalonyl pathway' is predicted by MetaCyc to function in *H. sapiens*. However, when viewing the 'glyoxylate and dicarboxylate metabolism' pathway across the phylogenetic profile shown in Figure 3.1, it seems to be relatively well conserved among bacteria. While the lack of orthologous proteins in *H. sapiens* might warrant further examination of these proteins as drug targets, their conservation in other bacteria might prove difficult in the actual application of drugs.

The reactions belonging to 'sulfur metabolism' also show few orthologs between Mtb and *H. sapiens.* Sulfur and sulfite can be used as sources of energy and electrons in organisms able to metabolise it (Kappler and Dahl, 2001). The bars representing 'sulfur metabolism' in Figure 3.3 show close to even numbers of orthologs in *C. glutamicum* and *E. coli.* This also shows very few shared orthologs in *H. sapiens.* Lastly, many of the proteins added for Mtb also have orthologs in *M. leprae.* The heat map of the

phylogenetic profile shown in Figure 3.1 shows that is pathway is relatively conserved across the genome. However, most of those proteins that were added to this pathway using BLASTP matches do not have orthologs across the phylogenetic profile. Most of these added proteins are involved in first the transport of taurine and alkanesulfonate from extracellular sources and second their conversions into sulfite. This would suggest that these proteins involved in transport and conversion of compounds that are precursors of sulfite might require further examination as drug targets, particularly since these proteins are not shared across the phylogenetic profile but have orthologs in the genome of *M. leprae*.

The 'geraniol degradation' pathway also shows few orthologs between Mtb and *H. sapiens.* In fact, Mtb proteins involved in this pathway share orthologies with very few proteins across the entire phylogenetic profile and the pathway is in the bottom third in terms of average frequency as shown in Figure 3.2. This pathway involves the degradation of citronellol, a compound produced by plants and found in the environment (Caspi et al., 2012). *Pseudomonas citronellolis* can use citronellol as a sole source of carbon and energy (Seubert, 1960). Mtb shows many more proteins assigned to this pathway than any of the organisms compared in Figure 3.3 suggesting that this pathway may be important for the organism. On the other hand, Mtb also shares very few orthologs for this pathway with *M. leprae*, indicating that perhaps these reactions are not essential for central metabolism. However, the much greater numbers of Mtb proteins belonging to this pathway, compared to other organisms, warrants investigation into the potential competitive advantages this pathway might provide within the host. Since citronellol is a common plant-derived compound in the environment, the existence of proteins catalysing reactions within this pathway may simply be an evolutionary artefact of a soil-dwelling ancestor. However, the large differences in numbers of orthologous proteins seen between Mtb and other organisms warrant further attention.

Lastly the 'polycyclic aromatic hydrocarbon degradation' pathway, along with many other pathways involved in 'xenobiotics biodegradation and metabolism', shows few orthologs between Mtb and *H. sapiens.* Additionally Mtb shares few orthologs with any other examined organism for this pathway (Figure 3.3); this pathway ranks in the

bottom 20% of pathways, signifying that it is one of the least conserved as shown in Figure 3.2. These results identify this pathway as being notable and worthy of further investigation. Many of the added proteins to this pathway are involved in the degradation of phenanthrene, a compound found in cigarette smoke. The most common way by which phenanthrene, along with many other polycyclic aromatic hydrocarbons (PAHs), enters the body is by breathing air contaminated with fumes from coal, asphalt, wildfires, vehicle exhaust or even grilled food (Phenanthrene). A number of species of mycobacteria have been shown to be able to grow on phenanthrene (and other PAHs) as a sole source of carbon and energy (Boldrin et al., 1993; Moody et al., 2001; Willumsen et al., 2001). Interestingly, exposure to tobacco has been shown to increase TB infection, active disease and mortality (Bates et al., 2007; Wen et al., 2010). This correlation is thought to occur because smoking suppresses the immune system, thus allowing greater chance of infection and progression of the disease as well as a reduced ability of the body to fight back against infection. However, these newly annotated proteins in Mtb offer another possibility: perhaps smoking increases virulence and survival of Mtb by providing a source of carbon and energy. The findings of the phylogenetic profile thus suggest that reactions within this pathway could represent potential drug targets that must be examined further.

An examination of the number of *H. sapiens* proteins that have orthologs in Mtb results in the identification of a number of interesting pathways. These pathways, including 'biotin metabolism', 'glyoxylate and dicarboxylate metabolism', 'sulfur metabolism', 'geraniol degradation' and 'polycyclic aromatic hydrocarbon degradation, signify pathways that could be targeted by novel drugs without adversely affecting human cells.

4.4 Missing Pathway Results

After annotating as many reactions as possible with proteins from Mtb H37Rv, some reactions remained uncharacterised. These 'missing' reactions were compiled into a table showing the pathway membership, reaction and evidence that shows they are missing (see Appendix C). This evidence includes the existence of the reaction in closely related organisms, its classification as 'missing' in Pathway Tools and its existence in the KEGG global map. Certain reactions have all three types of evidence supporting

their status as 'missing' and these can provide interesting cases. 'Missing' reactions or pathway holes result from one of two conditions. First enzymes catalysing these reactions may in fact exist but their sequence structure differs to a large degree from enzymes in closely related organisms. Secondly, these can be cases in which Mtb does not possess an enzyme catalysing this reaction and rather uses the host metabolism to accomplish these conversions.

One such reaction is for the interconversion of glycine and threonine. This is a central reaction within the 'glycine, serine and threonine metabolism' pathway that has been characterised in other related organisms and is classified as a pathway hole in Pathway Tools. It has been reported that a wide variety of bacteria have the ability to grow on L-threonine as a sole source of carbon and nitrogen (Bell and Turner, 1977). Additionally many organisms have been found to possess the L-threonine aldolase that catalyses this reaction (Morris, 1969). However, the physiological relevance of this enzyme in bacteria has not been established and many organisms do not exhibit high aldolase activity (Bell and Turner, 1977). Its absence from Mtb could mean that either it is not essential (due to the existence of alternative pathways between the two compounds) or the microbe utilises the host to achieve this interconversion.

A missing reaction belonging to the 'one carbon pool by folate' pathway converts 5,10-methylene-THF into 5-methyl-THF. This reaction is classified as a pathway hole in Pathway Tools and has been characterised in closely related organisms. The reaction is involved in both methionine biosynthesis and 1-carbon metabolism (Caspi et al., 2012). The enzyme catalysing this reaction, methylenetetrahydrofolate reductase, has been identified in *E. coli* as well as other bacteria and eukaryotes including humans (Sheppard et al., 1999). The widespread distribution of this enzyme suggests that it is very important for survival and growth. The inability to characterise this reaction in Mtb suggests that the enzyme either does not exist or has a different amino acid sequence compared to enzymes from closely related organisms. If the enzyme does not exist in Mtb it is possible that the microbe utilises the host metabolism to accomplish this conversion.

Another reaction that has been classified as a missing reaction in Pathway Tools, shows on the KEGG global map and has been characterised in closely related organisms

belongs to the 'folate biosynthesis' pathway. Very little information could be found regarding the enzyme catalysing this reaction, alkaline optimum 2-amino-4-hydroxy-6-(erythro-1,2,3-trihydroxypropyl) dihydropteridine triphosphate phosphohydrolase, but it catalyses the only missing reaction in the 'folate biosynthesis' pathway. Folates are essential cofactors of many reactions in which one-carbon units are transferred from donor molecules to important biosynthetic pathways involved in the biosynthesis of methionine, purine and pyrimidine, the conversion of histidine catabolism and interconversion of serine and glycine (Caspi et al., 2012; Lucock, 2000). Thus the absence of an enzyme catalysing this essential reaction in Mtb could be due to a divergent sequence for this enzyme in Mtb or to the utilisation of the host metabolism to accomplish this biosynthesis.

The reaction with the substrate 6-hexanolide and product 6-hydroxyhexanoate, catalysed by epsilon-caprolactone gluconolactonase (also 6-hexanolide hydrolase) and part of the 'caprolactam degradation' pathway, is also classified as 'missing' on the basis of all three types of evidence. Multiple other reactions were characterised for this pathway, and Figure 3.3 shows that many of these proteins have no orthologs in other organisms including *C. glutamicum, E. coli* and *H. sapiens.* The pathway is common within the phylum actinobacteria and typically the five enzymes converting cyclohexanol to adipate, including this 'missing' reaction function in one operon. Although the four other enzymes were characterised in Mtb using BLASTP matches, this reaction has yet to be found. It is thus likely that this enzyme exists in Mtb but simply has not been assigned a protein from the genome.

These few, as well as many other, reactions have been classified as 'missing' reactions or pathway holes and are shown in the appendix (Appendix C). These pathway holes deserve further attention to discover certain cases where Mtb may utilise the metabolome of the human host in order to accomplish its metabolic needs.

4.5 Conclusion

This study has identified a number of pathways and reactions that deserve further attention as potential drug targets. First, certain pathways for which many proteins have been assigned were investigated. Second, pathways showing interesting

characteristics across the phylogenetic profile were identified. Interesting pathways in this case include those in which Mtb shares both many and very few orthologs with *M. leprae,* many orthologs with *C. glutamicum,* and few orthologs with *E. coli* and *H. sapiens.* Lastly, the compiled table of 'missing' reactions or pathway holes was examined to find reactions which deserve further attention as possible cases where Mtb utilises the host genome to accomplish its metabolic requirements. These strategies have aided in the further characterisation of the genome of *M. tuberculosis* H37Rv.

5 Conclusion

Although many efforts have been made to reduce the burden of the disease, tuberculosis remains a major global health problem. An estimated 1.3 million people died of TB in 2012, including 88,000 people in South Africa alone (World Health Organization, 2013c, 2013f). In addition, the deadly combination of HIV and TB affects 1.1 million or the 8.6 million people infected with TB in 2012, with 75% of these instances occurring in sub-Saharan Africa. Lastly, the World Health Organization estimates that globally 3.6% of new cases and 20% of previously treated cases of TB are MDR-TB, with 9.6% of MDR-TB cases are actually XDR-TB (World Health Organization, 2013f). TB treatment and prevention urgently needs to be addressed in order to reduce this significant burden of the disease both around the world and in South Africa.

Several factors contribute to the high incidence and prevalence of TB around the world. One complication involved in the treatment of TB includes the lack of drugs affecting the bacteria during the persistent phase (Ma et al., 2010). Mtb persists in an early phagosome within macrophages where it resides in either a non-replicating or slowly replicating state, and no currently available drugs are effective against Mtb in this state (Kaufmann and Parida, 2008; Wakamoto et al., 2013). Secondly, poor diagnosis methods with low sensitivity such as sputum smear microscopy delay the treatment of TB and thus allow the disease to spread further (Abubakar et al., 2013; Jassal and Bishai, 2009). Third, the long treatment periods and many side effects of currently available drugs reduce patient compliance, causing increased risk of reactivation and development of drug resistance (Ma et al., 2010). HIV coinfection also makes treatment more difficult and carries with it an increased risk of reactivation of the disease (Jassal and Bishai, 2009; Ma et al., 2010). Lastly, drug resistance requires longer treatment periods with both first and second-line drugs (which have more side effects) and is more difficult to cure (Jassal and Bishai, 2009). All of these factors complicate the treatment of TB and underscore the need for new drugs.

New drugs are urgently needed to address these issues and improve treatment outcomes for patients. In order to efficiently and rationally design these new drugs, more research needs to be completed regarding the metabolism of Mtb. With further

understanding of its metabolism, drugs can be developed that target specific reactions or pathways in the metabolome. The identification and validation of appropriate targets is currently a bottleneck in the drug development pipeline. In fact, many drug in use today have been developed without specific drug targets in mind, probably because there are no standard methodologies to identify these targets on a large scale (Raman et al., 2008). These drug targets need to be essential for the survival of Mtb, preferably functional in Mtb during the persistent state and not have equivalent reactions or enzymes in *H. sapiens* and other beneficial bacteria that will be affected by the same drugs.

In order to find these new drug targets, this study first aimed to improve the characterisation of the Mtb metabolome, and then compared pathway and reaction information over the phylogenetic profile, which includes a large spectrum of organisms. Orthologs, or genes in different species with shared ancestry by vertical descent, were used to make this comparison across the various species by finding pathways showing different numbers of orthologs. This study has attempted to first further characterise the metabolism of Mtb, second compare its metabolism across the phylogenetic spectrum to find notable pathways that can be deemed potential drug targets, and third to classify pathway holes, or reactions without an assigned enzyme in the genome of Mtb.

The first step of this study involved the addition of functional annotations to the genome of Mtb. A combination of annotation methods using EC number, GO terms and BLASTP matches were used to help complete the metabolic map. A total of 553 proteins were added to pathways based on GO terms and EC numbers. In addition, 288 reactions were annotated within the metabolism of Mtb using BLASTP matches. Some pathways in particular had many newly annotated reactions and thus were examined further. Benzoate degradation, phenylalanine metabolism and glyoxylate and dicarboxylate metabolism displayed a relatively high number of added annotations and were mapped to KEGG diagrams to identify important newly characterised pathway sections. It is unclear why so many additional annotations could be made for *M. tuberculosis,* when the genome should already be fully annotated. One possible cause is that the annotation is simply out of date and needs to be updated for this organism. Another possibility is

based on the genomes used for annotation. Typical automatic annotation uses well-characterised genomes to generate functional annotations for newly sequenced genomes. The genomes used to locate BLASTP matches in this study were all closely related organisms, and it is possible that these genomes were annotated after Mtb, and thus not used to annotate the genome of Mtb as they were in the current study. Whatever the reason, a substantial number of additional genes were annotated using the methods described thus far, and the study highlights the need to keep annotation as current as possible.

The metabolic map with the novel functional annotations was then used to identify pathways and reactions of interest by comparing the number of orthologs per pathway for 392 organisms across the phylogenetic profile. Five characteristics of interesting pathways include Mtb having both many and very few orthologs with *M. leprae,* very few orthologs with *E. coli* and *H. sapiens* and many orthologs with *C. glutamicum*, a facultative anaerobe. Comparisons of the number of orthologs with *M. leprae,* an organism with a highly reduced genome, can show reactions that are likely to be essential in Mtb. Proteins of Mtb without orthologs in other organisms, such as *E. coli* and *H. sapiens* can classify enzymatic reactions, which, if attacked, would not affect the metabolism of host cells nor beneficial bacteria within the host. Lastly, by comparison with an anaerobic bacterium such as *C. glutamicum*, potential pathways essential for anaerobic metabolism can be identified, thereby finding potential drug targets for the persistent phase of infection. A number of interesting pathways were identified fulfilling these characteristics including arginine and proline metabolism and glycerolipid metabolism (many orthologs shared between Mtb and *M. leprae*), histidine metabolism and butanoate metabolism (few orthologs shared between Mtb and *E. coli*), xylene degradation and porphyrin and chlorophyll metabolism (many orthologs shared between Mtb and *C. glutamicum*) and lastly biotin metabolism, sulfur metabolism, geraniol degradation and polycyclic aromatic hydrocarbon degradation (few orthologs between Mtb and *H. sapiens*). Important functions of these pathways were then described to assess the essentiality of the pathway and significance during infection. These pathways have therefore been identified as potential drug targets for the development of novel drugs.

The third and final goal of this project was to identify pathway holes, or essential reactions either without annotated enzymes or instances in which Mtb might use the host metabolism to accomplish its metabolic needs. Pathway holes were located by mapping the enzymes of Mtb onto KEGG pathway diagrams. Reactions without annotated enzymes in the genome of Mtb were cross-referenced with pathway holes identified by Pathway Tools and reactions that have annotated enzymes in the genomes of closely related organisms but for which no BLASTP matches could be found. A total of 363 pathway holes were classified, which can help elucidate the interaction between host and microbe during infection.

The use of sequence similarity as a proxy for functional similarity, while being supported in the literature, also has its limitations. Some proteins with highly similar sequences may not share function, especially if they have experienced duplication events and horizontal gene transfer such as has occurred in Mtb (Kelley et al., 2003). The use of EC number and GO terms to annotate the genome of Mtb relies on the accuracy of these classifications and this also limits the accuracy of the annotations. Lastly, the use of KEGG as the reference database can have a major effect since functional characterisation of the metabolome relies on the precision of the database. Repetitious reactions and the lack of taxonomic predictors for pathway function in different organisms can cause false positives in terms of pathway membership and pathway hole prediction. Additionally since KEGG pathway diagrams include all reactions pertaining to a particular topic, comparison of the number of proteins in M. tuberculosis H37Rv for each pathway with the number of orthologous proteins for organisms across the phylogenetic profile might not always be effective. For example, two organisms might show the same number of orthologous proteins but have entirely different pathway branches within the diagram, and this must be taken into account.

In summary, many new annotations were added to the genome of Mtb and a number of pathways were identified as potential drug targets for future drug development. These annotations should be added to the genome of Mtb in the KEGG database by updating the genome annotation. Future research should be directed toward those pathways identified as potential drug targets. Enzymes within these identified pathways must be investigated further to evaluate this potential. A combination of expression data (to

discover activity during the persistent phase), protein-protein interactions and assessments of essentiality can be used with knowledge of currently available compound specificity to develop novel drugs that will help reduce the burden that is caused by tuberculosis globally.

References

Abubakar, I., Zignol, M., Falzon, D., Raviglione, M., Ditiu, L., Masham, S., Adetifa, I., Ford, N., Cox, H., Lawn, S.D., Marais, B.J., McHugh, T.D., Mwaba, P., Bates, M., Lipman, M., Zijenah, L., Logan, S., McNerney, R., Zumla, A., Sarda, K., Nahid, P., Hoelscher, M., Pletschette, M., Memish, Z.A., Kim, P., Hafner, R., Cole, S., Migliori, G.B., Maeurer, M., Schito, M., Zumla, A., 2013. Drug-resistant tuberculosis: time for visionary political leadership. Lancet Infect. Dis. 13, 529–39.

Alam, M.T., Merlo, M.E., Takano, E., Breitling, R., 2010. Genome-based phylogenetic analysis of Streptomyces and its relatives. Mol. Phylogenet. Evol. 54, 763–72.

Alber, B.E., Spanheimer, R., Ebenau-Jehle, C., Fuchs, G., 2006. Study of an alternate glyoxylate cycle for acetate assimilation by Rhodobacter sphaeroides. Mol. Microbiol. 61, 297–309.

Allison, M., Daniel, S., Cornick, N., 1995. Oxalate-degrading bacteria, in: Khan, S.R. (Ed.), Calcium Oxalate in Biological Systems. CRC Press, Inc., Boca Raton, Florida, pp. 131–168.

Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., Dessimoz, C., 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput. Biol. 8, e1002514.

Altman, T., Travers, M., Kothari, A., Caspi, R., Karp, P.D., 2013. Open Access A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics 14, 112.

Aoki-Kinoshita, K.F., Kanehisa, M., 2007. Gene annotation and pathway mapping in KEGG. Methods Mol. Biol. 396, 71–91.

Bates, M.N., Khalakdina, A., Pai, M., Chang, L., Lessa, F., Smith, K.R., 2007. Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis. Arch. Intern. Med. 167, 335–42.

Baxevanis, A.D., Ouellette, B.F.F., 2005. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Google eBook). John Wiley & Sons, Hoboken, New Jersey.

Beckloff, N., Starkenburg, S., Freitas, T., Chain, P., 2012. Bacterial Genome Annotation. Methods Mol. Biol. 881, 471–503.

Bell, S.C., Turner, J.M., 1977. Bacterial catabolism of threonine. Threonine degradation initiated by L-threonine acetaldehyde-lyase (aldolase) in species of Pseudomonas. Biochem. J. 166, 209–16.

Bentrup, K.H. zu, Russell, D.G., 2001. Mycobacterial persistence: adaptation to a changing environment. Trends Microbiol. 9, 597–605.

Beste, D.J.V., McFadden, J., 2013. Metabolism of Mycobacterium tuberculosis, in: McFadden, J., Beste, D.J.V., Kierzek, A.M. (Eds.), Systems Biology of Tuberculosis. Springer New York, New York, NY, pp. 55–78.

Boerjan, W., Ralph, J., Baucher, M., 2003. Lignin biosynthesis. Annu. Rev. Plant Biol. 54, 519–46.

Boldrin, B., Tiehm, A., Fritzsche, C., 1993. Degradation of phenanthrene, fluorene, fluoranthene, and pyrene by a Mycobacterium sp. Appl. Environ. Microbiol. 59, 1927–1930.

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J., Zaretskaya, I., 2013. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 41, W29–33.

Borek, B.A., Waelsch, H., 1953. The enzymatic degradation of histidine. J. Biol. Chem. 205, 459–74.

Brennan, M.J., Thole, J., 2012. Tuberculosis vaccines: a strategic blueprint for the next decade. Tuberculosis (Edinb). 92 Suppl 1, S6–13.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., Karp, P.D., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 42, D459–71.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P., Karp, P.D., 2012. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 40, D742–D753.

Caspi, R., Dreher, K., Karp, P.D., 2013. The challenge of constructing, classifying, and representing metabolic pathways. FEMS Microbiol. Lett. 345, 85–93.

Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P., Karp, P.D., 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 34, D511–6.

Chen, X., Zhang, J., 2012. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing. PLoS Comput. Biol. 8.

Clark, D.P., Cronan, J.E., 2005. Two-Carbon Compounds and Fatty Acids as Carbon Sources. EcoSal Plus.

Colditz, G.A., Brewer, T.F., Berkey, C.S., Wilson, M.E., Burdick, E., Fineberg, H.V., Mosteller, F., 1994. Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. J. Am. Med. Assoc. 271, 698–702.

Daniel, T.M., 2006. The history of tuberculosis. Respir. Med. 100, 1862–1870.

De Carvalho, L.P.S., Fischer, S.M., Marrero, J., Nathan, C., Ehrt, S., Rhee, K.Y., 2010. Metabolomics of Mycobacterium tuberculosis reveals compartmentalized co-catabolism of carbon substrates. Chem. Biol. 17, 1122–31.

Dhar, N., McKinney, J.D., 2010. Mycobacterium tuberculosis persistence mutants identified by screening in isoniazid-treated mice. Proc. Natl. Acad. Sci. U. S. A. 107, 12275–80.

Dye, C., Harries, A.D., Maher, D., Hosseini, S.M., Nkhoma, W., Salaniponi, F.M., 2006. Tuberculosis, in: Jamison, D.T., Feachem, R.G., Makgoba, M.W., Bos, E.R., Baingana, F.K., Hofman, K.J., Rogo, K.O. (Eds.), Disease and Mortality in Sub-Saharan Africa. World Bank, Washington DC.

Ederveen, T.H.A., Overmars, L., van Hijum, S.A.F.T., 2013. Reduce manual curation by combining gene predictions from multiple annotation engines, a case study of start codon prediction. PLoS One 8, e63523.

Erb, T.J., Berg, I.A., Brecht, V., Müller, M., Fuchs, G., Alber, B.E., 2007. Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: the ethylmalonyl-CoA pathway. Proc. Natl. Acad. Sci. U. S. A. 104, 10631–6.

Evans, W.C., 1963. the Microbiological Degradation of Aromatic Compounds. J. Gen. Microbiol. 32, 177–84.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J.P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R., Searle, S.M.J., 2014. Ensembl 2014. Nucleic Acids Res. 42, D749–55.

Francois, A., Mathis, H., Godefroy, D., Piveteau, P., Fayolle, F., Monot, F., 2002. Biodegradation of Methyl tert-Butyl Ether and Other Fuel Oxygenates by a New Strain, Mycobacterium austroafricanum IFP 2012. Appl. Environ. Microbiol. 68, 2754–2762.

Fuchs, G., 2008. Anaerobic metabolism of aromatic compounds. Ann. N. Y. Acad. Sci. 1125, 82–99.

Galagan, J.E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., Gomes, A., Rustad, T., Dolganov, G., Glotova, I., Abeel, T., Mahwinney, C., Kennedy, A.D., Allard, R., Brabant, W., Krueger, A., Jaini, S., Honda, B., Yu, W.-H., Hickey, M.J., Zucker, J., Garay, C., Weiner, B., Sisk, P., Stolte, C., Winkler, J.K., Van de Peer, Y., Iazzetti, P., Camacho, D., Dreyfuss, J., Liu, Y., Dorhoi, A., Mollenkopf, H.-J., Drogaris, P., Lamontagne, J., Zhou, Y., Piquenot, J., Park, S.T., Raman, S., Kaufmann, S.H.E., Mohney, R.P., Chelsky, D., Moody, D.B., Sherman, D.R., Schoolnik, G.K., 2013. The Mycobacterium tuberculosis regulatory network and hypoxia. Nature 499, 178–83.

Garcia-Betancur, J.C., Menendez, M.C., Del Portillo, P., Garcia, M.J., 2012. Alignment of multiple complete genomes suggests that gene rearrangements may contribute towards the speciation of Mycobacteria. Infect. Genet. Evol. 12, 819–26.

Getahun, H., Harrington, M., O'Brien, R., Nunn, P., 2007. Diagnosis of smear-negative pulmonary tuberculosis in people with HIV infection or AIDS in resource-constrained settings: informing urgent policy changes. Lancet 369, 2042–9.

Ginsburg, B., Lovett, S.L., Dunn, M.S., 1956. Amino acid composition of mycobacteria. Arch. Biochem. Biophys. 60, 164–170.

Gopinath, K., Moosa, A., Mizrahi, V., Warner, D.F., 2013. Vitamin B-12 metabolism in Mycobacterium tuberculosis. Future Microbiol. 8, 1405–1418.

Green, M.L., Karp, P.D., 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics 5, 76.

Guirado, E., Schlesinger, L.S., Kaplan, G., 2013. Macrophages in tuberculosis: friend or foe. Semin. Immunopathol. 35, 563–83.

Gupta, A., 2009. Annotation, in: Liu, L., Özsu, M.T. (Eds.), Encyclopedia of Database Systems. Springer, pp. 85–85.

Gutierrez, M.C., Brisse, S., Brosch, R., Fabre, M., Omaïs, B., Marmiesse, M., Supply, P., Vincent, V., 2005. Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. PLoS Pathog. 1, 55–61.

Hartmans, S., Bont, J., Stackebrandt, E., 2006. The Genus Mycobacterium- Nonmedical, in: Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H., Stackebrandt, E. (Eds.), Prokaryotes Volume 3: Archaea. Bacteria: Firmicutes Actinomycetes. pp. 889–918.

Harwood, C.S., Burchhardt, G., Herrmann, H., Fuchs, G., 1998. Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway. FEMS Microbiol. Rev. 22, 439–458.

Harwood, C.S., Parales, R.E., 1996. The beta-ketoadipate pathway and the biology of self-identity. Annu. Rev. Microbiol. 50, 553–90.

Heider, J., Fuchs, G., 1997. Anaerobic Metabolism of Aromatic Compounds. Eur. J. Biochem. 243, 577–596.

Hochuli, M., Patzelt, H., Oesterhelt, D., Szyperski, T., 1999. Amino Acid Biosynthesis in the Halophilic Archaeon Haloarcula hispanica Amino Acid Biosynthesis in the Halophilic Archaeon Haloarcula hispanica. J. Bacteriol. 181, 3226–3237.

Hong, S.H., Kim, T.Y., Lee, S.., 2004. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl. Microbiol. Biotechnol. 65, 203–10.

Jassal, M., Bishai, W.R., 2009. Extensively drug-resistant tuberculosis. Lancet Infect. Dis. 9, 19–30.

Jenkinson, C.P., Grody, W.W., Cederbaum, S.D., 1996. Comparative properties of arginases. Comp. Biochem. Physiol. B. Biochem. Mol. Biol. 114, 107–32.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L., 2008. NCBI BLAST: a better web interface. Nucleic Acids Res. 36, W5–9.

Kalinowski, J., Bathe, B., Bartels, D., Bischoff, N., Bott, M., Burkovski, A., Dusch, N., Eggeling, L., Eikmanns, B.J., Gaigalat, L., Goesmann, A., Hartmann, M., Huthmacher, K., Krämer, R., Linke, B., McHardy, A.C., Meyer, F., Möckel, B., Pfefferle, W., Pühler, A., Rey, D.A., Rückert, C., Rupp, O., Sahm, H., Wendisch, V.F., Wiegräbe, I., Tauch, A., 2003. The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of l-aspartate-derived amino acids and vitamins. J. Biotechnol. 104, 5–25.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y., 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480–4.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M., 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355–60.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–7.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32, D277–80.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–14.

Kappler, U., Dahl, C., 2001. Enzymology and molecular biology of prokaryotic sulfite oxidation. FEMS Microbiol. Lett. 203, 1–9.

Karp, P.D., Latendresse, M., Caspi, R., 2011. The pathway tools pathway prediction algorithm. Stand. Genomic Sci. 5, 424–9.

Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., Caspi, R., 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief. Bioinform. 11, 40–79.

Karp, P.D., Riley, M., Paley, S.M., Pellegrini-Toole, A., 2002. The MetaCyc Database. Nucleic Acids Res. 30, 59–61.

Kaufmann, S.H.E., Hussey, G., Lambert, P.-H., 2010. New vaccines for tuberculosis. Lancet 375, 2110–9.

Kaufmann, S.H.E., Parida, S.K., 2008. Tuberculosis in Africa: learning from pathogenesis for biomarker identification. Cell Host Microbe 4, 219–28.

Kaufmann, S.H.E., van Helden, P., 2008. Handbook of Tuberculosis: Clinics, Diagnostics, Therapy and Epidemiology. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., Ideker, T., 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. U. S. A. 100, 11394–9.

Kerfeld, C.A., Scott, K.M., 2011. Using BLAST to teach "E-value-tionary" concepts. PLoS Biol. 9.

Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I., Apweiler, R., 2005. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res. 33, D297–302.

Koch, R., Brock, T.D., Fred, E.B., 1982. The Etiology of Tuberculosis, 1882. Rev. Infect. Dis. 4, 1270–1274.

Koonin, E. V., 2005. Orthologs, Paralogs, and Evolutionary Genetics. Annu. Rev. Genet. 39, 309–338.

LeBlanc, J.C., Gonçalves, E.R., Mohn, W.W., 2008. Global response to desiccation stress in the soil actinomycete Rhodococcus jostii RHA1. Appl. Environ. Microbiol. 74, 2627–36.

Lehmann, J., 1946. Para-Aminosalicylic Acid In The Treatment of Tuberculosis. Lancet 247, 15–16.

Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., Corbett, M., Cheng, Y., Demiralp, F., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Hunter, C., Jang, M., Leonard, S., Lin, Q., Lopez, R., Maguire, M., McWilliam, H., Plaister, S., Radhakrishnan, R., Sobhany, S., Slater, G., Ten Hoopen, P., Valentin, F., Vaughan, R., Zalunin, V., Zerbino, D., Cochrane, G., 2010. Improvements to services at the European Nucleotide Archive. Nucleic Acids Res. 38, D39–45.

Lerner, K.L., Lerner, B.W., 2008. Sequencing. Gale Encycl. Sci.

Lilley, B.D., Brewer, J.H., 1953. The selective antibacterial action of phenylethyl alcohol. J. Am. Pharm. Assoc. 42, 6–8.

Lin, S., Hanson, R.E., Cronan, J.E., 2010. Biotin synthesis begins by hijacking the fatty acid synthetic pathway. Nat. Chem. Biol. 6, 682–8.

Lucock, M., 2000. Folic acid: nutritional biochemistry, molecular biology, and role in disease processes. Mol. Genet. Metab. 71, 121–38.

Luengo, J.M., Garcia, J.L., Olivera, E.R., 2001. The phenylacetyl-CoA catabolon: a complex catabolic unit with broad biotechnological applications. Mol. Microbiol. 39, 1434–1442.

Lunardi, J., Eduardo S. Nunes, J., V. Bizarro, C., Augusto Basso, L.A., Santiago Santos, D., Machado, P., 2013. Targeting the Histidine Pathway in Mycobacterium tuberculosis. Curr. Top. Med. Chem. 13, 2866–2884.

Ma, Z., Lienhardt, C., McIlleron, H., Nunn, A.J., Wang, X., 2010. Global tuberculosis drug development pipeline: the need and the reality. Lancet 375, 2100–9.

Mao, X., Chen, X., Zhang, Y., Pangle, S., Xu, Y., 2012. CINPER: an interactive web system for pathway prediction for prokaryotes. PLoS One 7, e51252.

Marri, P.R., Bannantine, J.P., Golding, G.B., 2006. Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer. FEMS Microbiol. Rev. 30, 906–25.

Martens, J.H., Barg, H., Warren, M.J., Jahn, D., 2002. Microbial production of vitamin B12. Appl. Microbiol. Biotechnol. 58, 275–85.

Matsunaga, I., Bhatt, A., Young, D.C., Cheng, T.-Y., Eyles, S.J., Besra, G.S., Briken, V., Porcelli, S.A., Costello, C.E., Jacobs, W.R., Moody, D.B., 2004. Mycobacterium tuberculosis pks12 produces a novel polyketide presented by CD1c to T cells. J. Exp. Med. 200, 1559–69.

Moat, A.G., Foster, J.W., Spector, M.P., 2003. Central Pathways of Carbohydrate Metabolism, in: Microbial Physiology. John Wiley & Sons, Inc., pp. 350–367.

Moody, J.D., Freeman, J.P., Doerge, D.R., Cerniglia, C.E., 2001. Degradation of phenanthrene and anthracene by cell suspensions of Mycobacterium sp. strain PYR-1. Appl. Environ. Microbiol. 67, 1476–83.

Moore, B.S., Hertweck, C., Hopke, J.N., Izumikawa, M., Kalaitzis, J.A., Nilsen, G., O'Hare, T., Piel, J., Shipley, P.R., Xiang, L., Austin, M.B., Noel, J.P., 2002. Plant-like biosynthetic pathways in bacteria: from benzoic acid to chalcone. J. Nat. Prod. 65, 1956–62.

Morgat, A., Coissac, E., Coudert, E., Axelsen, K.B., Keller, G., Bairoch, A., Bridge, A., Bougueleret, L., Xenarios, I., Viari, A., 2012. UniPathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Res. 40, D761–9.

Morris, J.G., 1969. Utilization of L-threnonine by a pseudomonad: a catabolic role for L-threonine aldolase. Biochem. J. 115, 603–5.

Mulder, N.J., Kersey, P., Pruess, M., Apweiler, R., 2008. In silico characterization of proteins: UniProt, InterPro and Integr8. Mol. Biotechnol. 38, 165–77.

Munoz-Elias, E.J., McKinney, J.D., 2005. Mycobacterium tuberculosis isocitrate lyases 1 and 2 are jointly required for in vivo growth and virulence. Nat. Med. 11, 638–644.

National Institute for Health and Clinical Excellence, 2011. Tuberculosis: Clinical diagnosis and management of tuberculosis, and measures for its prevention and control. London.

Nehrt, N.L., Clark, W.T., Radivojac, P., Hahn, M.W., 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput. Biol. 7, e1002073.

Nishimura, T., Vertès, A.A., Shinoda, Y., Inui, M., Yukawa, H., 2007. Anaerobic growth of Corynebacterium glutamicum using nitrate as a terminal electron acceptor. Appl. Microbiol. Biotechnol. 75, 889–97.

O'Callaghan, D., Stebbins, C.E., 2010. Host-microbe interactions: bacteria. Curr. Opin. Microbiol. 13, 1–3.

Oberhardt, M.A., Palsson, B.O., Papin, J.A., 2009. Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. 5, 320.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M., 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29–34.

Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., Palsson, B.O., 2003. Metabolic pathways in the post-genome era. Trends Biochem. Sci. 28, 250–8.

Park, S.W., Hwang, E.H., Park, H., Kim, J.A., Heo, J., Lee, K.H., Song, T., Kim, E., Ro, Y.T., Kim, S.W., Kim, Y.M., 2003. Growth of Mycobacteria on Carbon Monoxide and Methanol. J. Bacteriol. 185, 142–147.

Parrott, S., Jones, S., Cooper, R.A., 1987. 2-Phenylethylamine catabolism by Escherichia coli K12. J. Gen. Microbiol. 133, 347–51.

Patterson, J., Waller, R., Jeevarajah, R., Billman-Jacobe, H., McConville, M., 2003. Mannose metabolism is required for mycobacterial growth. Biochem. J. 372, 77–86.

Petryszak, R., Kretschmann, E., Wieser, D., Apweiler, R., 2005. The predictive power of the CluSTr database. Bioinformatics 21, 3604–9.

Phenanthrene, n.d. Atlanta, Georgia.

Pieters, J., 2008. Mycobacterium tuberculosis and the macrophage: maintaining a balance. Cell Host Microbe 3, 399–407.

Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E., Ziv-Ukelson, M., 2005. Alignment of metabolic pathways. Bioinformatics 21, 3401–8.

Pirovano, W., Heringa, J., 2008. Multiple Sequence Alignment, in: Keith, J.M. (Ed.), Bioinformatics, Volume I: Data, Sequence Analysis, and Evolution, Methods in Molecular Biology Vol. 452. Humana Press, Totowa, NJ, pp. 143–162.

Raman, K., Yeturu, K., Chandra, N., 2008. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC Syst. Biol. 2, 109.

Reddy, A.S., Zhang, S., 2013. Polypharmacology: drug discovery for the future. Expert Rev. Clin. Pharmacol. 6, 41–7.

Rhee, K.Y., de Carvalho, L.P.S., Bryk, R., Ehrt, S., Marrero, J., Park, S.W., Schnappinger, D., Venugopal, A., Nathan, C., 2011. Central carbon metabolism in Mycobacterium tuberculosis: an unexpected frontier. Trends Microbiol. 19, 307–14.

Richardson, E.J., Watson, M., 2013. The automatic annotation of bacterial genomes. Brief. Bioinform. 14, 1–12.

Rush, D., Karibian, D., Karnovsky, M.L., Magasanik, B., 1957. Pathways of glycerol dissimilation in two strains of Aerobacter aerogenes; enzymatic and tracer studies. J. Biol. Chem. 226, 891–9.

Saeed, A.I., Bhagabati, N.K., Braisted, J.C., Liang, W., Sharov, V., Howe, E.A., Li, J., Thiagarajan, M., White, J.A., Quackenbush, J., 2006. TM4 microarray software suite. Methods Enzymol. 411, 134–93.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., Quackenbush, J., 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34, 374–8.

Saka, H.A., Valdivia, R.H., 2010. Acquisition of nutrients by Chlamydiae: unique challenges of living in an intracellular compartment. Curr. Opin. Microbiol. 13, 4–10.

Sakula, A., 1983. Robert koch: centenary of the discovery of the tubercle bacillus, 1882. Can. Vet. J. 24, 127–31.

Sawers, G., 1998. The anaerobic degradation of L-serine and L-threonine in enterobacteria: networks of pathways and regulatory signals. Arch. Microbiol. 171, 1–5.

Schatz, A., 2005. Streptomycin, a substance exhibiting antibiotic activity against Gram-positive and Gram-negative bacteria. Clin. Orthop. Relat. Res. 437, 3.

Schilling, C.H., Letscher, D., Palsson, B.O., 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J. Theor. Biol. 203, 229–48.

Schuhle, K., Gescher, J., Feil, U., Paul, M., Jahn, M., Schagger, H., Fuchs, G., 2003. Benzoate-Coenzyme A Ligase from Thauera aromatica: an Enzyme Acting in Anaerobic and Aerobic Pathways. J. Bacteriol. 185, 4920–4929.

Sekine, M., Tanikawa, S., Omata, S., Saito, M., Fujisawa, T., Tsukatani, N., Tajima, T., Sekigawa, T., Kosugi, H., Matsuo, Y., Nishiko, R., Imamura, K., Ito, M., Narita, H., Tago, S., Fujita, N., Harayama, S., 2006. Sequence analysis of three plasmids harboured in Rhodococcus erythropolis strain PR4. Environ. Microbiol. 8, 334–46.

Seubert, W., 1960. Degradation of isoprenoid compounds by micro-organisms. I. Isolation and characterization of an isoprenoid-degrading bacterium, Pseudomonas citronellolis n. sp. J. Bacteriol. 79, 426–34.

Sheppard, C.A., Trimmer, E.E., Matthews, R.G., 1999. Purification and properties of NADH-dependent 5, 10-methylenetetrahydrofolate reductase (MetF) from Escherichia coli. J. Bacteriol. 181, 718–25.

Shiloh, M.U., Champion, P.A.D., 2010. To catch a killer. What can mycobacterial models teach us about Mycobacterium tuberculosis pathogenesis? Curr. Opin. Microbiol. 13, 86–92.

Smith, L.T., Pocard, J.A., Bernard, T., Le Rudulier, D., 1988. Osmotic control of glycine betaine biosynthesis and degradation in Rhizobium meliloti. J. Bacteriol. 170, 3142–9.

Smith, N.H., Hewinson, R.G., Kremer, K., Brosch, R., Gordon, S. V., 2009. Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis. Microbiology 7, 537–544.

Stadthagen, G., Korduláková, J., Griffin, R., Constant, P., Bottová, I., Barilone, N., Gicquel, B., Daffé, M., Jackson, M., 2005. p-Hydroxybenzoic acid synthesis in Mycobacterium tuberculosis. J. Biol. Chem. 280, 40699–706.

Sukrasno, N., Yeoman, M.M., 1993. Phenylpropanoid metabolism during growth and development of Capsicum frutescens fruits. Phytochemistry 32, 839–844.

Tameris, M.D., Hatherill, M., Landry, B.S., Scriba, T.J., Snowden, M.A., 2013. Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled phase 2b trial - ProQuest. Lancet 381, 1021–1028.

Teufel, R., Mascaraque, V., Ismail, W., Voss, M., Perera, J., Eisenreich, W., Haehnel, W., Fuchs, G., 2010. Bacterial phenylalanine and phenylacetate catabolic pathway revealed. Proc. Natl. Acad. Sci. U. S. A. 107, 14390–5.

The Gene Ontology Consortium, 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–9.

The Gene Ontology Consortium, 2013. Gene Ontology annotations and resources. Nucleic Acids Res. 41, D530–5.

The UniProt Consortium, 2014. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 42, D191–D198.

Thomas, P.D., Wood, V., Mungall, C.J., Lewis, S.E., Blake, J.A., 2012. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. PLoS Comput. Biol. 8, e1002386.

Torres, O.H., Domingo, P., Pericas, R., Boiron, P., Montiel, J.A., Vázquez, G., 2000. Infection Caused by Nocardia farcinica : Case Report and Review. Eur. J. Clin. Microbiol. Infect. Dis. 19, 205–212.

Tortoli, E., 2012. Phylogeny of the genus Mycobacterium: many doubts, few certainties. Infect. Genet. Evol. 12, 827–31.

Trunz, B.B., Fine, P., Dye, C., 2006. Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. Lancet 367, 1173–80.

Tummler, B., 2010. Genome Annotation, in: Timmis, K.N. (Ed.), Handbook of Hydrocarbon and Lipid Microbiology. Springer-Verlag Berlin Heidelberg, pp. 4281–4288.

US National Institute of Allergy and Infectious Disease, 2009. Tuberculosis (TB).

Valderrama, J.A., Durante-Rodríguez, G., Blázquez, B., García, J.L., Carmona, M., Díaz, E., 2012. Bacterial degradation of benzoate: cross-regulation between aerobic and anaerobic pathways. J. Biol. Chem. 287, 10494–508.

Van der Werf, M.J., Swarts, H.J., de Bont, J.A., 1999. Rhodococcus erythropolis DCL14 contains a novel degradation pathway for limonene. Appl. Environ. Microbiol. 65, 2092–102.

Van Ingen, J., Boeree, M.J., van Soolingen, D., Iseman, M.D., Heifets, L.B., Daley, C.L., 2012. Are phylogenetic position, virulence, drug susceptibility and in vivo response to treatment in mycobacteria interrelated? Infect. Genet. Evol. 12, 832–7.

Vicuña, R., 1988. Bacterial degradation of lignin. Enzyme Microb. Technol. 10, 646–655.

Wakamoto, Y., Dhar, N., Chait, R., Schneider, K., Signorino-Gelo, F., Leibler, S., McKinney, J.D., 2013. Dynamic persistence of antibiotic-stressed mycobacteria. Science 339, 91–5.

Wang, X., Wang, H., Xie, J., 2011. Genes and Regulatory Networks Involved in Persistence of Mycobacterium tuberculosis. Sci. China 54, 300–310.

Warren, M.J., Raux, E., Schubert, H.L., Escalante-Semerena, J.C., 2002. The biosynthesis of adenosylcobalamin (vitamin B12). Nat. Prod. Rep. 19, 390–412.

Wen, C.-P., Chan, T.-C., Chan, H.-T., Tsai, M.-K., Cheng, T.-Y., Tsai, S.-P., 2010. The reduction of tuberculosis risks by smoking cessation. BMC Infect. Dis. 10, 156.

Williams, P.A., Murray, K., 1974. Metabolism of Benzoate and the Methylbenzoates by mt-2 : Evidence for the Existence of a TOL Plasmid Metabolism of Benzoate and the Methylbenzoates by Pseudomonas putida (arvilla) mt-2 : Evidence for the Existence of a TOL Plasmid. J. Bacteriol. 120, 416–423.

Willumsen, P.A., Nielsen, J.K., Karlson, U., 2001. Degradation of phenanthrene-analogue azaarenes by Mycobacterium gilvum strain LB307T under aerobic conditions. Appl. Microbiol. Biotechnol. 56, 539–544.

World Health Organization, 2011. The Global Plan to Stop TB 2011 – 2015.

World Health Organization, 2012a. Global Tuberculosis Report 2012.

World Health Organization, 2012b. 2011/2012 Tuberculosis Global Facts.

World Health Organization, 2013a. Multidrug-resistant tuberculosis (MDR-TB).

World Health Organization, 2013b. HIV-Associated TB Facts 2013.

World Health Organization, 2013c. South Africa Tuberculosis profile.

World Health Organization, 2013d. Tuberculosis Diagnostics.

World Health Organization, 2013e. Tuberculosis Fact Sheet No. 104, WHO. World Health Organization.

World Health Organization, 2013f. Global Tuberculosis Report 2013.

Wu, G., 2009. Amino acids: metabolism, functions, and nutrition. Amino Acids 37, 1–17.

Yang, X., Gao, J., Smith, I., Dubnau, E., Sampson, N.S., 2011. Cholesterol is not an essential source of nutrition for Mycobacterium tuberculosis during infection. J. Bacteriol. 193, 1473–6.

Ye, J., McGinnis, S., Madden, T.L., 2006. BLAST: improvements for better sequence analysis. Nucleic Acids Res. 34, W6–9.

Young, D.B., Gideon, H.P., Wilkinson, R.J., 2009. Eliminating latent tuberculosis. Trends Microbiol. 17, 183–8.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.-D.J., Bertin, N., Chung, S., Vidal, M., Gerstein, M., 2004. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res. 14, 1107–18.

Zahrt, T.C., 2003. Molecular mechanisms regulating persistent Mycobacterium tuberculosis infection. Microbes Infect. 5, 159–167.

Zeller, E.A., Van Orden, L.S., Vogtli, W., 1954. Enzymology of mycobacteria. VII. Degradation of guanidine derivatives. J. Biol. Chem. 209, 429–35.

Zhang, L., Zhang, C., Cheng, Z., Yao, Y., Chen, J., 2013. Biodegradation of benzene, toluene, ethylbenzene, and o-xylene by the bacterium Mycobacterium cosmeticum byf-4. Chemosphere 90, 1340–7.

Zhang, Y., 2005. The magic bullets and tuberculosis drug targets. Annu. Rev. Pharmacol. Toxicol. 45, 529–64.

Zhi, X.-Y., Li, W.-J., Stackebrandt, E., 2009. An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class Actinobacteria, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. Int. J. Syst. Evol. Microbiol. 59, 589–608.

Appendix A

Proteins Annotated with BLASTP Matches, Remaining Pathways

Blue= not previously annotated in pathways, Red= 'uncharacterised' proteins, Yellow= other strains

Pathway	# of Prot in Path	New Proteins Found	# Missing
Pyrimidine Metabolism	49	P0A616 & P65548 (GO); O53217 (GO); O53590 (GO); P68911 (Msmeg); L0TED1/O05791 (Msmeg); P71809 (Msmeg)	5
Purine Metabolism	79	P65548 (GO); O07732 (GO); O50399 (GO); O33230 (GO)(dup); O53707 (GO); O53699 (GO); Q7D4M2 (GO); C6DMB2 (GO); A5U160 (GO); P68911 (Msmeg); P71809 (Strep); O53772 & Q11058 (Mvan)	9 (maybe 8)
Tyrosine Metabolism	58	P0A678 (GO); O53904 & P95153 (GO); O53303 & O53533 & Q7DAC8 & P71818 & O53904 & P0A4X0 & P95153 & O07737 (Msmeg); O86346 (Rjost); Q8VK36 (Rjost); P95275 (Rjost); P71865 (Rjost); O53242 & O86346 (Rjost); P63937 & P96405 & O33340 & P71823 & O53816 & P71989 & P96417 & Q7D5R7 & P96824 & O50443 (Nfarc); C6DW56 (GO)(KZN)	13
Tryptophan Metabolism	67	P63492 (GO); O06544 (GO); Q11146 (Rjost); P77900 & O08447 & P63719 & O69653 & P0A512 & P63717 & P95099 & P63715 & O33180 (Nfarc); P63516 (Rjost); P96405 & P96417 & P71989 & O53816 & Q7D5R7 & O50443 (Nfarc); O50463 (Strep); O53929 (Rery); O07205 (Rery)	14
Phenylalanine, Tyrosine and Tryptophan Biosynthesis	20	P96257 (Msmeg)	0
Cysteine and Methionine Metabolism	29	P96847? (GO); P66875 & P95199 & O53390 (Msmeg); P95075 (Nfarc)	8
Valine, Leucine and Isoleucine Degradation	83	P96855 & O53577 & P95097 (GO); O06160 & O06161 (Msmeg); O06335 & P71867 (Msmeg); Q8VK36 (Msmeg); P63427 & O06164 & O33229 & O33331 & O53815 & P95208 & P96397 & O86319 & P71539 & P63429 & P95187 & P96844 & P95281 & P96855 & O53549 & P71858 & O53926 & P95097 & P96831 & P95228 (Rjost); P95280 & O53577 & P96842 (Rjost)	4
Valine, Leucine and Isoleucine Biosynthesis	14	None	1
Lysine Biosynthesis	16	P63509 (EC); O53407 (GO); O53176 & P95139 (Msmeg)	11
Lysine Degradation	50	Q7D529 (GO); O53176 & P95139 (Msmeg); O05820 (Rjost); O50463 (Strep)	3
Alanine, Aspartate and	32	P96847 (GO); O53446 (Msmeg); Q10896 & L0TB18 & O05819 (Rjost); O69644 (Rery)	1

Glutamate Metabolism			
Steroid Biosynthesis	1		Unknown
Fatty Acid Metabolism	70	Q10878 & Q7D5D8 & O53551 & O07797 (GO/EC); O06544 (GO); O07737 (EC); O05893 (GO); O06164 & P95186 & O86319 & P63429 & P95208 & P96842 & O33331 & P96831 & P96397 (Msmeg); O33229 & O86319 & O06164 & P63427 & O53815 & P96397 & P71539 & O33331 & P95208 & O53549 & P71858 & P95280 & P63429 & P95186 & P96831 & P96855 & P96808 & O53577 (Msmeg); P77900 & O08447 & P63719 & O69653 & P0A512 & P63717 & P95099 & P63715 & O33180 (Nfarc); P95034 & P95146 (Mvan); P63427 & P96397 & O06164 & O53815 & O33229 & P95208 & P63429 & O33331 & O86319 & P96831 & P71539 & P95208 & O53549 & P95186 & P96808 & O53577 & P96844 & P71858 & P96842 & P95228 & O53926 (Rjost)	0
Synthesis and Degradation of Ketone Bodies	41	O06334 & P71867 (Msmeg); P69167 & P95286 & O33339 & P0A5Y4 & P71824 & P95101 & P95273 & O33292 & P66781 & P96841 & O53665 & O53398 & Q11150 & O50460 & P71853 & P71871 & P96825 & Q7D6M3 & P95033 & O53547 & O06348 & O53927 & O33263 & O50417 & O53302 & O05919 & O06413 (Msmeg)	2
Alpha-Linolenic Acid Metabolism	11	O53567 (EC); O06164 & P95186 & O86319 & P63429 & P95208 & P96842 & O33331 & P96831 & P96397 (Msmeg)	7/9
Biosynthesis of Unsaturated Fatty Acids	22	P63640 (GO); O06164 & P95186 & O86319 & P63429 & P95208 & P96842 & O33331 & P96831 & P96397 (Msmeg)	11
Glycerophospholipid Metabolism	23	O07427 & P0A642 (GO)	11
Ether Lipid Metabolism	7	None	Unknown
Fatty Acid Biosynthesis	33	P71980 (EC); P0A5Y4 (EC); P0A574 (Msmeg); P63456 & P63454 & O53579 & Q10977 & O86335 & P96284 & P94996 & L0TA10 & O53901 & P96202 & P96291 & L0T5X5 & P71718 & O06586 & O65933 & P96204 & O07798 (Msmeg); P0A5Y6 & P71871 & O05919 & P71824 (Strep)	0
Linoleic Acid Metabolism	5	None	Unknown
Sphingolipid Metabolism	4	None	Unknown
Steroid Biosynthesis	1	None	Unknown
Glycolysis/ Gluconeogenesis	42	O07737 (EC); A5U5J3 & C6DMY5 (EC)(CDC & KZN); P0A4X0 (EC)	3
Citrate Cycle (TCA Cycle)	37	O53671 (GO); O50463 (EC); O50463 (Strep)	0

Pyruvate Metabolism	48	P64261 (EC); P96886 (EC); P65684 & O06579 (Msmeg); P95040 (Msmeg); P0A622 & P66946 & O06335 & O53639 & O53554 (Msmeg)	3
Propanoate Metabolism	80	O86318 & O53578 & O06165 & P63407 (EC); P66984 (Msmeg); O06831 & P95227 & O07411 & P96396 & O53306 & P96843 & P96283 & P71716 & Q7D5D8 & O06417 & O53406 & O06168 & Q10878 & Q50586 (Msmeg); Q8VK36 (Msmeg); P0A5H3 & O07718 (Msmeg)	9
Pentose Phosphate Pathway	20	P71825 (EC); P65154 (Msmeg)	6
Pentose and Glucuronate Interconversions	44	P95286 & P0A5Y4 & O53927 & P95033 & P66781 & P71824 & O33339 & P95273 & O33292 & P69167 & P66777 & O06348 & O05919 & P96841 & O33263 & P95101 & P71852 & O06544 & P71564 & P71853 & O53863 & P71871 & P96825 & O53665 & Q7D6M3 & O53547 & O53302 & O50460 & P95150 & O53398 (Msmeg); O69664 (Msmeg); P71821 & Q11150 & O50417 (Msmeg); O06413 (Msmeg); P95075 (Msmeg)	5 (or 13)
Ascorbate and Aldarate Metabolism	7	P95075 (Msmeg)	6
Starch and Sucrose Metabolism	28	O07242 (GO); P71741 (EC)	4
Amino Sugar and Nucleotide Sugar Metabolism	38	P64905 (GO); Q50685 (EC); P63338 (Msmeg); P95277 & Q7ARS9 & O05875 & P96853 & P71846 & O86347 (Msmeg)	4
C5-Branched Dibasic Acid Metabolism	10	P0A666 (Msmeg)	1
Galactose Metabolism	14	P67475 (Msmeg)	0
Inositol Phosphate Metabolism	16	Q7D6W6 (GO); P96283 & P66946 & O53639 & O53865 (Msmeg); O86352 (Msmeg)	4
Oxidative Phosphorylation	49	P64947 (EC); O53671 (GO)	0
Nitrogen Metabolism	46	P71753 (EC); P63627 (EC); O06179 (EC); A5TYS2 & P71994 (EC); Q7D4Q3 (EC); P96218 (Msmeg); Q11146 (Rjost)	3
Methane Metabolism	57	P64118 & P65408 (EC); A5UBZ1 (EC); A5TZ90 (EC); P65573 & P65567 & P95175 & P65408 & O06817 & O06560 & L0T2Z1 (Msmeg); O53294 & P64745 & P71662 & O53762 & P96223 & O53300 (Msmeg); P96253 & Q10814 & P96809 & O07914 & P95140 & P95159 & O53565 & P71701 & P64769 (Msmeg); P65684 & O06579 (Msmeg); O53533 (UniPathway); O07737 & O53904 & Q7DAC8 & P72043 & P0A4X0 & P95153 (Strep)	16
Polyketide Sugar Unit Biosynthesis	6	None	0

Biosynthesis of Siderophore Group Nonribosomal Peptides	9	None	0
Terpenoid Backbone Biosynthesis	27	None	0
Carotenoid Biosynthesis	1	None	Unknown
Ubiquinone and Other Terpenoid-Quinone Biosynthesis	21	P96843 & O06168 & O07411 & O53306 & O53406 & P95227 & O06417 & O05295 & P96396 & O07169 & O06831 & O53551 & O53521 & Q7D5D8 (Rjost); Q8VK36 (Rjost)	3
Thiamine Metabolism	10	None	2
Riboflavin Metabolism	11	P95275 (GO)	1
Vitamin B6 Metabolism	11	O53240 & A5WM18 & O07171 & O50398 & O06553 (EC); P66913 (Msmeg)	4
Nicotinate and Nicotinamide Metabolism	18	P96394 (EC); P68911 (Msmeg)	2
Pantothenate and CoA Biosynthesis	19	P71809 (Msmeg)	3 (or 5)
Lipoic Acid Metabolism	2	None	Unknown
Peptidoglycan Biosynthesis	27	P71707 (GO); O33346 (Strep)	3
Lipopolysaccharide Biosynthesis	6	P95231 (Rjost)	2 (or 9)
Taurine and Hypotaurine Metabolism	21	O53379 & P63504 & P63568 & P0A4X6 & P63506 & P63509 & P71890 & P71891 (Msmeg); Q50685 & Q11061 & O07406 & L0TBR2 & P72056 (Rjost)	3
Beta-Alanine Metabolism	48	P71809 (Msmeg); Q8VKI0 (Msmeg); O53379 & P63504 & P63568 & P0A4X6 & P63506 & P63509 & P71890 & P71891 (Msmeg)	6
Selenocompound Metabolism	12	P95199 & O53390 (Msmeg)	1
D-Glutamine and D-Glutamate Metabolism	4	O69644 (Rjost)	1
D-Arginine and D-	1	None	2

Ornithine Metabolism			
Cyanoamino Acid Metabolism	7	Q11146 (Rjost)	1 (could be many more)
D-Alanine Metabolism	2	None	0
Glutathione Metabolism	18	O05915 & O53356 (EC); P0A5M4 (Nfarc)	3
Penicillin and Cephalosporin Biosynthesis	2	None	Unknown
Novobiocin Biosynthesis	3	None	Unknown
Streptomycin Biosynthesis	11	None	3 (but could be more)
Fluorobenzoate Degradation	37	P95277 & P66781 & P71871 & O33339 & P71824 & P96841 & Q7ARS9 & P69167 & O33292 & P95286 & O05919 & Q11150 & O53398 & P95033 & P96853 & P95273 & O33263 & Q7D6M3 & P71846 & O53863 & P0A5Y4 & P95101 & O06413 & P71853 & P71821 (Msmeg); O53665 & P71852 & O53927 & P96825 & O50417 & O50460 & O53302 (Rjost); P65425 (Rjost); O65936 & O53772 (Rjost); P95118 (Rjost); P65083 (Rjost)	0
Chlorocyclohexane and Chlorobenzene Degradation	35	P96850 (Msmeg); P65425 (Rjost); O50405 (Msmeg); P96811 & O06266 & P0A572 & P95276 & O06576 (Msmeg); P95034 & P95146 (Mvan); O53311 & O53674 & P66006 & O07927 & O06598 & O53641 & P96839 (Rjost); O65936 & O53772 (Rjost); P95118 (Rjost); P66777 & O69638 & O53321 & O06420 (Rjost); P65083 (Rjost); O05301 & O06339 (Rjost)	2
Nitrotoluene Degradation	6	O50431 & O50388 (Mvan)	2
Styrene Degradation	21	Q11146 (Rjost); P96850 & O33319 & O86347 (Rjost); O86346 (Nfarc); P71850 (Nfarc); O33340 & P96405 & P96417 & P63937 & P71823 & P71989 & O53816 & Q7D5R7 & P96824 & O50443 (Strep)	4 (or 7)
Atrazine Degradation	11	O69719 (GO); P0A676 & P0A660 & P0A662 (EC); O53258 & P63490 & P63494 & P63496 & O53325 & P63492 (Msmeg); P71809 (Rjost)	4 (or 2)
Naphthalene Degradation	69	P96825 & P71853 & O53547 & P66781 & P71824 & P95033 & O33292 & O53863 & P96841 & O53398 & O50417 & P0A5Y4 & O53665 & O53302 & P69167 & O50460 & P95101 & P66777 & P95273 & P95286 & P71821 & O33263 & O33339 & O05919 & Q7D6M3 & P71871 & P71852 & O53927 & P66779 & O06348 & P95150 & O06413 & O53533 & O69693 & O53146 & P0A4X0 & P72043 & P95185 & O05842 & O53613 & O53726 & O53537 & O07230 & O53324 & Q11150 & Q10782 & P96202 (Msmeg);	4

		O53772 & Q11058 (Rjost/Strep); O53311 & P95034 & P95146 & O53674 & P66006 & O07927 & O06598 & O53641 & P96839 (Rjost)	
Steroid Degradation	14	Q50616 & O53670 (Msmeg)	0
Aminobenzoate Degradation	108	O53580 (EC); Q11146 (Rjost); O06335 & P0A622 & O53639 & P66946 & O53554 (Msmeg); P96405 & P71823 & O33340 & P96417 & P71989 & P63937 & O53816 & Q7D5R7 & P96824 & O50443 (Msmeg); P95034 & P95146 & O53674 & Q79FW1 (Msmeg); O65936 (Cglut); P71825 & O86347 & P96853 & P71846 & O05875 (Cglut); P96283 & O07411 & O05295 & O53306 & O06831 & P95227 & O69635 & O06168 & P96396 & O06417 & O53406 & P96843 & P71716 & O07169 & Q7D5D8 & Q10878 (Nfarc); O50388 & O50431 (Nfarc); P77900 & O08447 & P63719 & O69653 & P0A512 & P63717 & P95099 & P63715 & O33180 (Nfarc); O53555 (Rjost); P95034 & P95146 (Mvan)	3
Chloroalkane and Chloroalkene Degradation	68	O07737 (EC); P96825 & P71853 & O53547 & P66781 & P71824 & P95033 & O33292 & O53863 & P96841 & O53398 & O50417 & P0A5Y4 & O53665 & O53302 & P69167 & O50460 & P95101 & P66777 & P95273 & P95286 & P71821 & O33263 & O33339 & O05919 & Q7D6M3 & P71871 & P71852 & O53927 & P66779 & O06348 & P95150 & O06413 & O69693 & O53146 & P95185 & O05842 & O53613 & O53726 & O53537 & O07230 & O53324 & Q11150 & Q10782 & P96202 (Msmeg); O06266 & P0A572 & P95276 & O06576 (Msmeg); O50405 (Msmeg); P95034 & P95146 (Mvan); O53533 & P72043 & P0A4X0 (Strep)	4
Toluene Degradation	36	O53671 (EC); O53303 & O53533 & Q7DAC8 & P71818 & O53904 & P0A4X0 & P95153 & O07737 (Msmeg); P95118 (Rjost); P65425 (Rjost); P65083 (Rjost); P95034 & P95146 (Mvan); O65936 (Cglut); Q50685 (Rjost); O53311 & O53674 & P66006 & O07927 & O06598 & O53641 & P96839 (Rjost)	7
Ethylbenzene Degradation	24	O53567 (EC); P95034 & P95146 & O53674 & Q79FW1 (Msmeg) P96850 (Rjost); P96851 & P0A572 & P64303 & O06266 & O53327 & P95276 & P64301 & P96811 & O69638 & O06420 (Rjost)	1
Bisphenol Degradation	89	P96825 & P71853 & O53547 & P66781 & P71824 & P95033 & O33292 & O53863 & P96841 & O53398 & O50417 & P0A5Y4 & O53665 & O53302 & P69167 & O50460 & P95101 & P66777 & P95273 & P95286 & P71821 & O33263 & O33339 & O05919 (Msmeg); Q7D6M3 & P71871 & P71852 & O53927 & P66779 & O06348 & P95150 & O06413 & O07737 & O53303 & O53533 & O69693 & P71818 & Q7DAC8 & O53904 & P95153 & O53146 & P0A4X0 & P72043 & P95185 & O05842 & O53613 & O53726 & O53537 & O07230 & O53324 & Q11150 & Q10782 & P96202 (Msmeg); O53294 & P71662 & P64745 & P96223 & O53762 & O53300 (Mvan)	2
Dioxin Degradation	31	P96851 (GO)(confirmed with Rjost); P96850 (Msmeg); P71865 (Mvan); P95034 & P95146 (Mvan); O53300 & P71824 & P95033 & P66781 & O33292 & O05919 & Q11150 & P69167 & P71871 & P71853 & Q7D6M3 & O53863 & P95273 & O05842 & P0A5Y4 & P71852 & O33339 & P96841 & P95150 & O53302 & O06544 (Rjost); O53772 & Q11058 (Rjost)	0

Appendix B

List of Organisms from Figure 3.1

1. M.tuberculosis Oshkosh
2. M.tuberculosis KZN 1435
3. M.tuberculosis F11
4. M.tuberculosis ATCC 25618
5. M.tuberculosis ATCC 25177
6. M.bovis Tokyo 172
7. M.bovis Pasteur 1173P2
8. M.bovis AF2122/97
9. Mycobacterium sp. MCS
10. Mycobacterium sp. KMS
11. Mycobacterium sp. JLS
12. M.vanbaalenii
13. M.ulcerans
14. M.smegmatis
15. M.paratuberculosis
16. M.leprae TN
17. M.leprae Br4923
18. M.gilvum
19. M.avium
20. Rhodococcus sp.
21. R.erythropolis
22. N.farcinica
23. C.jeikeium
24. C.glutamicum Nakagawa
25. C.glutamicum Kalinowski
26. C.efficiens
27. C.diphtheriae
28. T.whipplei Twist
29. T.fusca
30. S.tropica
31. S.coelicolor
32. S.avermitilis
33. P.acnes DSM 16379
34. Nocardioides sp.
35. L.xyli
36. K.radiotolerans
37. Frankia sp. EAN1pec
38. Frankia sp. CcI3
39. F.alni
40. Arthrobacter sp.
41. A.cellulolyticus
42. B.longum NCC 2705
43. R.xylanophilus
44. Z.mobilis CP4
45. Y.pseudotuberculosis I
46. Y.pestis 91001
47. X.oryzae MAFF 311018
48. X.fastidiosa 9a5c
49. X.campestris vesicatoria
50. X.axonopodis
51. X.autotrophicus
52. Wolbachia sp. Brugia malayi
53. W.succinogenes
54. W.pipientis wMel
55. V.vulnificus YJ016
56. V.parahaemolyticus
57. V.fischeri ATCC 700601
58. V.eiseniae
59. V.cholerae ATCC 39315
60. Thermoanaerobacter sp. X514
61. T.thermophilus HB8
62. T.tengcongensis
63. T.pseudethanolicus
64. T.maritima
65. T.erythraeum
66. T.elongatus
67. T.denticola
68. T.denitrificans
69. T.crunogena
70. Synechocystis sp.
71. Synechococcus sp. CC9311
72. Silicibacter sp.
73. Shewanella sp. MR-4
74. S.wolfei
75. S.usitatus
76. S.typhimurium ATCC 700720
77. S.typhi ATCC 700931
78. S.thermophilum
79. S.sonnei
80. S.schwarzengrund
81. S.saprophyticus
82. S.ruber DSM 13855
83. S.pyogenes MGAS5005
84. S.putrefaciens CN-32
85. S.pomeroyi
86. S.pneumoniae ATCC BAA-255
87. S.paratyphi SARB42
88. S.oneidensis
89. S.newport
90. S.mutans ATCC 700610
91. S.medicae
92. S.loihica
93. S.heidelberg
94. S.haemolyticus
95. S.glossinidius
96. S.gallinarum
97. S.fumaroxidans
98. S.frigidimarina
99. S.flexneri 5b
100. S.epidermidis ATCC 35984
101. S.enteritidis
102. S.elongatus
103. S.dysenteriae
104. S.dublin
105. S.denitrificans OS217
106. S.degradans
107. S.choleraesuis
108. S.boydii 4
109. S.baltica OS195
110. S.aureus Mu50
111. S.amazonensis
112. S.alaskensis
113. S.agona
114. S.agalactiae Ia
115. S.aciditrophicus
116. Roseiflexus sp.

117. R.sphaeroides ATCC 17023
118. R.solanacearum
119. R.rubrum
120. R.palustris HaA2
121. R.metallidurans
122. R.meliloti
123. R.loti
124. R.leguminosarum bv. viciae
125. R.ferrireducens
126. R.etli CFN 42
127. R.denitrificans
128. R.baltica
129. Psychrobacter sp.
130. Polynucleobacter sp.
131. Polaromonas sp.
132. P.vibrioformis
133. P.ubique
134. P.thermopropionicum
135. P.syringae tomato
136. P.putida KT2440
137. P.propionicus
138. P.profundum
139. P.phaeoclathratiforme
140. P.pentosaceus
141. P.multocida
142. P.marinus MIT 9313
143. P.luteolum
144. P.luminescens
145. P.ingrahamii
146. P.haloplanktis
147. P.gingivalis W83
148. P.fluorescens Pf-5
149. P.entomophila
150. P.denitrificans
151. P.cryohalolentis
152. P.carbinolicus
153. P.atlantica
154. P.arcticus
155. P.amoebophila
156. P.aestuarii
157. P.aeruginosa LMG 12228
158. O.oeni
159. O.iheyensis
160. Nostoc sp.
161. N.winogradskyi
162. N.oceani
163. N.multiformis
164. N.meningitidis
165. N.hamburgensis
166. N.gonorrhoeae ATCC 700825
167. N.eutropha
168. N.europaea
169. N.aromaticivorans
170. Mesorhizobium sp.
171. Magnetococcus sp.
172. M.xanthus
173. M.thermoacetica
174. M.succiniciproducens
175. M.magneticum
176. M.capsulatus
177. L.welshimeri
178. L.sakei
179. L.reuteri JCM 1112
180. L.reuteri DSM 20016
181. L.pneumophila Paris
182. L.plantarum WCFS1
183. L.plantarum JDM1
184. L.monocytogenes EGD-e
185. L.mesenteroides
186. L.lactis IL1403
187. L.lactis
188. L.johnsonii NCC 533
189. L.intracellularis
190. L.interrogans copenhageni
191. L.innocua
192. L.delbrueckii ATCC BAA-365
193. L.casei ATCC 334
194. L.brevis
195. L.borgpetersenii L550
196. L.acidophilus
197. Jannaschia sp.
198. I.loihiensis
199. H.somnus 129Pt
200. H.pylori 26695
201. H.orenii
202. H.neptunium
203. H.influenzae ATCC 51907
204. H.hepaticus
205. H.halophila
206. H.chejuensis
207. H.aurantiacus
208. Geobacter sp. FRC-32
209. G.violaceus
210. G.uraniireducens
211. G.sulfurreducens ATCC 51573
212. G.oxydans
213. G.metallireducens
214. G.kaustophilus
215. G.bethesdensis
216. F.tularensis LVS
217. F.nucleatum nucleatum
218. F.johnsoniae
219. E.sibiricum
220. E.ruminantium CIRAD
221. E.litoralis
222. E.faecalis V583
223. E.coli MG1655
224. E.chaffeensis
225. E.carotovora
226. E.canis
227. Dehalococcoides sp. CBDB1
228. D.vulgaris DP4
229. D.reducens
230. D.radiodurans
231. D.psychrophila
232. D.hafniense Y51
233. D.geothermalis
234. D.ethenogenes
235. D.desulfuricans G20
236. D.aromatica
237. Caulobacter sp.
238. C.violaceum
239. C.trachomatis D/UW-3/Cx
240. C.thermocellum ATCC 27405
241. C.tetani
242. C.tepidum
243. C.salexigens
244. C.saccharolyticus
245. C.psychrerythraea
246. C.pneumoniae J138

247. C.pinatubonensis
248. C.phytofermentans
249. C.phaeobacteroides DSM 266
250. C.perfringens 13
251. C.necator
252. C.muridarum
253. C.limicola
254. C.jejuni O:2
255. C.hydrogenoformans
256. C.hutchinsonii
257. C.difficile 630
258. C.crescentus CB15
259. C.chlorochromatii
260. C.burnetii Nine Mile phase
261. C.beijerinckii
262. C.aurantiacus ATCC 29366
263. C.acetobutylicum
264. Burkholderia sp. ATCC 17760
265. Bradyrhizobium sp. BTAi1
266. B.xenovorans
267. B.weihenstephanensis
268. B.vietnamiensis
269. B.thuringiensis konkukian
270. B.thetaiotaomicron
271. B.thailandensis
272. B.suis 1
273. B.subtilis 168
274. B.quintana
275. B.pseudomallei K96243
276. B.pertussis
277. B.parapertussis
278. B.ovis
279. B.microti
280. B.melitensis 1
281. B.mallei ATCC 23344
282. B.licheniformis Novozymes
283. B.japonicum
284. B.henselae
285. B.halodurans
286. B.fragilis ATCC 25285
287. B.floridanus
288. B.clausii
289. B.cicadellinicola
290. B.cereus ZK
291. B.cenocepacia HI2424
292. B.canis
293. B.bronchiseptica
294. B.bacteriovorus
295. B.avium
296. B.anthracis Sterne
297. B.ambifaria AMMD
298. B.abortus 2308
299. Acinetobacter sp. ADP1
300. Acidovorax sp.
301. A.vinelandii
302. A.variabilis
303. A.tumefaciens
304. A.succinogenes
305. A.phagocytophilum
306. A.oremlandii
307. A.metalliredigens
308. A.marginale St. Maries
309. A.ehrlichei
310. A.dehalogenans 2CP-C
311. A.cryptum
312. A.citrulli
313. A.borkumensis
314. A.bacterium
315. A.aromaticum
316. A.aeolicus
317. Y.lipolytica
318. U.methanogenic
319. U.maydis
320. T.volcanium
321. T.nigroviridis
322. T.cruzi
323. T.acidophilum
324. S.tokodaii
325. S.solfataricus ATCC 35092
326. S.pombe
327. S.acidocaldarius
328. R.norvegicus
329. P.torridus
330. P.nodorum
331. P.furiosus
332. P.aerophilum
333. P.abyssi
334. O.tauri
335. O.sativa
336. N.pharaonis
337. N.fumigata ATCC MYA-4609
338. N.crassa
339. M.thermoautotrophicum
340. M.stadtmanae
341. M.oryzae
342. M.musculus
343. M.mazei
344. M.marisnigri
345. M.maripaludis S2
346. M.kandleri
347. M.jannaschii
348. M.hungatei
349. M.burtonii
350. M.barkeri
351. M.acetivorans
352. H.walsbyi
353. H.sapiens
354. H.salinarium
355. H.marismortui
356. G.gallus
357. D.rerio
358. D.pseudoobscura
359. D.melanogaster
360. D.discoideum
361. C.neoformans B-3501A
362. C.globosum NBRC 6347
363. C.elegans
364. C.briggsae
365. B.taurus
366. A.thaliana
367. A.terreus
368. A.pernix
369. A.oryzae
370. A.fulgidus
371. A.aegypti

Appendix C

Pathway Holes Table

Pathway L3	Reaction	EC #	Pathway Tools	Missing in KEGG Global	In other orgs
Pyrimidine Metabolism	Thymine <-> dihydrothymine AND uracil <-> dihydrouracil	1.3.1.1/2	Yes	Yes	No
	3-ureidoisobutyrate -> 3-amino-isobutanoate AND 3-ureidopropionate -> beta-alanine	3.5.1.6	No	Yes	No
	Cytidine <-> cytosine	2.4.2.2 3.2.2.8	No	No	No
	Deoxycytidine -> dCMP	2.7.1.74	No	No	No
	Deoxyuridine <-> uracil	2.4.2.23	No	No	No
Purine Metabolism	IDP -> IMP	3.6.1.5/6	No	No	No
	AMP -> IMP	3.5.4.6	No	No	No
	Adenine -> Hypoxanthine	3.5.4.2	No	No	No
	5-hydroxyisourate -> 5-hydroxy-2-oxo-4-ureido-2,5-dihydro-1H-imidazole-5-carboxylate	3.5.2.17	No	Yes	Yes
	5-hydroxy-2-oxo-4-ureido-2,5-dihydro-1H-imidazole-5-carboxylate -> (S)-allantoin	4.1.1.-	No	Yes	Yes
	(R)-allantoin <-> (S)-allantoin	5.1.99.3	No	Yes	Yes
	Allantoate <-> urea	3.5.3.4	No	Yes	Yes
	3',5'-cyclic AMP -> AMP	3.1.4.17/53	No	No	No
	Inosine <-> IMP? AND GMP <-> guanosine?	2.7.1.73?	No	No	No
Histidine Metabolism	L-Histidine -> urocanate	4.3.1.3	No	Yes	Yes
	Urocanate -> 4-imidazolone-5-propanoate	4.2.1.49	No	Yes	Yes
	N-formimino-L-glutamate -> L-glutamate	3.5.3.8	No	Yes	No
Tyrosine Metabolism	Tyrosine -> L-DOPA	1.10.3.1 1.14.18.1 1.14.16.2	No	No	No
	L-DOPA -> Dopamine	4.1.1.25 4.1.1.28	No	Yes	No
	Dopamine -> L-noradrenaline	1.14.17.1	No	Yes	No
	L-noradrenaline -> L-adrenaline	2.1.1.28	No	Yes	No
	Tyrosine -> 3-iodo-tyrosine AND 3-iodo-tyrosine -> 3,5-diiodo-tyrosine AND 3,5-diiodo-tyrosine -> triiodothronine	1.11.1.8	No	Yes	No
	4-hydroxyphenylpyruvate -> 4-	4.1.1.80	No	No	No

	hydroxy-phenylacetaldehyde				
	Homogentisate -> 4-maleyl-acetoacetate	1.13.11.5	No	Yes	Yes
	4-maleyl-acetoacetate -> 4-fumaryl-acetoacetate	5.2.1.2	No	Yes	No
	3,4-dihydroxy-phenylacetaldehyde <-> 3,4-dihydroxy-phenylacetate AND 3-methoxy-4-hydroxyphenyl-acetaldehyde -> homovanillate AND 3,4-dihydroxy-mandeladehyde <-> 3,4-dihydroxy-mandelate AND 3-methoxy-4-hydroxy-phenylglycol-aldehyde -> 3-methoxy-4-hydroxy-mandelate	1.2.1.5	No	No	No
	Homoprotocatechuate -> 2-hydroxy-5-carboxy-methylmuconate semialdehyde	1.13.11.15	No	No	No
	5-carboxymethyl-2-hydroxymuconate -> 5-carboxy-2-oxohept-3-enedioate	5.3.3.10	No	No	No
	5-carboxy-2-oxohept-3-enedioate -> 2-hydroxyhepta-2,4-dienedioate	4.1.1.68	No	No	No
	2,4-dihydroxyhept-2-enedioate -> succinate semialdehyde	HpaI	No	No	Yes
Tryptophan Metabolism	Tryptophan → 5-hydroxytryptophan	1.14.16.4	No	No	No
	5-hydroxytryptophan → serotonin AND tryptophan → tryptamine	4.1.1.28	No	Yes	No
	Serotonin → N-acetylserotonin	2.3.1.87	No	Yes	No
	N-acetylserotonin → melatonin	2.1.1.4	No	Yes	No
	Tryptophan → indolepyruvate	2.6.1.27	Yes	No	No
	Tryptophan → indole-3-acetamide	1.13.12.3	No	No	No
	indole-3-acetonitrile → indole-3-acetamide	4.2.1.84	No	No	Yes
	Indoleacetate → 2-formamino-benzoylacetate	1.13.11.-	No	No	No
	Tryptophan → N-formyl-kynurenine	1.13.11.11/52	No	No	No
	Anthranilate → 3-hydroxyanthranilate	1.14.16.3	No	Yes	No
	L-kynurenine → 3-hydroxy-L-kynurenine	1.14.13.9	No	Yes	No

	3-hydroxyanthranilate → 2-amino-3-carbocymuconate semialdehyde	1.13.11.6	No	Yes	No
	2-amino-3-carbocymuconate semialdehyde → 2-aminomuconate semialdehyde	4.1.1.45	No	Yes	No
	2-aminomuconate → 2-oxoadipate	1.5.1.-	No	No	No
Phenylalanine Metabolism	Phenylalanine -> trans-cinnamate	4.3.1.24 4.3.1.25	No	Yes	No
	Trans-cinnamate -> trans-4-hydroxycinnamate	1.14.13.11	No	Yes	No
	4-coumaroyl-coA -> caffeoyl-coA	1.14.13.-	No	Yes	No
	feruloyl-coA -> 4-hydroxy-3-methoxyphenyl-beta-hydroxypropionyl-coA	4.2.1.101	No	Yes	No
	4-hydroxy-3-methoxyphenyl-beta-hydroxypropionyl-coA -> vanillin	4.1.2.41	No	Yes	No
	Phenylalanine -> phenyl-ethylamine	4.1.1.28 4.1.1.53	No	No	No
	Trans-2,3-dihydroxy-cinnamate -> 2-hydroxy-6-oxononatrienedioate AND 2,3-dihydroxy-phenylpropanoate -> 2-hydroxy-6-oxonona-2,4-diene-1,9-dioate	1.13.11.16	No	Yes	Yes
Arginine and Proline Metabolism	Arginine -> ornithine + urea	3.5.3.1	Yes	Yes	No
	Urea-1-carboxylate -> CO2	3.5.1.54	No	Yes	Yes
	NH3 <-> carbamoyl-P	2.7.2.2	Yes	Yes	No
	NH3 -> carbamoyl-P	6.3.4.16	Yes	Yes	No
	L-proline -> D-proline	5.1.1.4	No	No	No
	L-proline -> trans-4-hydroxy-L-proline	1.14.11.2	No	No	No
	trans-4-hydroxy-L-proline -> L-1-proline-3-hydroxy-5-carboxylate	PRODH2	No	No	No
	D-4-hydroxy-2-oxoglutarate -> glyoxylate	4.1.3.16 4.1.1.3	No	No	Yes
	Trans-4-hydroxy-L-proline <-> cis-4-hydroxy-D-proline	5.1.1.8	No	No	No
	Arginine <-> guanidinoacetate	2.1.4.1	No	No	No
	Guanidinoacetate -> creatine	2.1.1.2	No	No	No
	Creatine <-> creatinine	3.5.2.10	No	No	Yes
	Creatinine -> N-methyl-hydantoin	3.5.4.1	No	No	Yes
	N-methyl-hydantoin -> N-carbamoyl-sarcosine	3.5.2.14	No	No	Yes
	N-carbamoyl-sarcosine -> sarcosine	3.5.1.59	No	No	No
	Agmatine -> N-carbamoyl-putrescine	3.5.3.12	Yes	No	Yes
	Agmatine -> putrescine	3.5.3.11	Yes	No	Yes
	N4-aceyl-aminobutanoate -> 4-aminobutanoate	3.5.1.63	No	No	No
	S-adenosyl-L-methionine	4.1.1.50	Yes	No	No
Cysteine and Methionine	L-homoserine -> O-succinyl-L-homoserine	2.3.1.46	Yes	Yes	No

Metabolism					
	S-adenosyl-L-methionine -> S-adenosyl-methioninamine	4.1.1.50	Yes	Yes	No
	S-methyl-5-thio-D-ribose -> S-methyl-5-thio-D-ribose 1-phosphate	2.7.1.100	No	Yes	No
	2,3-diketo-5-methyl-thiopentyl-1-phosphate -> 1,2-dihydroxy-3-keto-5-methyl-thiopentene	3.1.3.77	No	Yes	No but in others
	1,2-dihydroxy-3-keto-5-methyl-thiopentene -> 4-methylthio-2-oxobutanoate	1.13.11.54	No	Yes	No
	4-methylthio-2-oxobutanoate -> L-methionine	2.6.1.5 2.6.1.57	No	Yes	No
	S-adenosyl-L-methionine -> S-adenosyl-L-homocysteine	2.1.1.37	Yes	Yes	No
	S-D-ribosyl-L-homocysteine -> L-homocysteine	4.4.1.21	No	No	No
Valine, Leucine and Isoleucine Degradation	4-methyl-2-oxopentanoate -> 3-methylbutanoyl-coA AND 3-methyl-2-oxobutanoate -> isobutyryl-coA AND 3-methyl-2-oxopentanoate -> (S)-2-methyl-butanoyl-coA	1.2.7.7	Yes	No	No
	S-(3-methyl-butanoyl)-dihydrolipoamide-E -> 3-methylbutanoyl-coA + dihydro-lipoamide-E AND S-(2-methyl-propanoyl)-dihydrolipoamide-E -> isobutyryl-coA + dihydrolipoamide-E AND S-(2-methyl-butanoyl)-dihydrolipoamide-E -> (S)-2-methyl-butanoyl-coA + dihydro-lipoamide-E	2.3.1.168	No	Yes	No
	3-methyl-glutaconyl-coA <-> (S)-3-hydroxy-3-methylglutaryl-coA	4.2.1.18	No	Yes	No
	(S)-3-hydroxy-isobutyryl-coA -> (S)-3-hydroxy-isobutyrate	3.1.2.4	Yes	No	No
Valine, Leucine and Isoleucine Biosynthesis	Pyruvate + acetyl-coA -> (R)-2-methylmalate	2.3.1.182	No	Yes	No
Lysine Biosynthesis	L-2-amino-adipate -> 5-adenyl-2-aminoadipate -> Alpha-aminoadipoyl-S-acyle enzyme -> L-2-aminoadipate 6-semialdehyde	1.2.1.31	No	Yes	No
	L-2-aminoadipate -> LysW-gamma-L-alpha-aminoadipate	LysX	No	Yes	No
	LysW-gamma-L-alpha-aminoadipate -> LysW-gamma-L-alpha-aminoadipyl 6-phosphate	LysZ	No	Yes	No
	LysW-gamma-L-alpha-aminoadipyl 6-phosphate -> LysW-gamma-L-alpha-aminoadipate 6-semialdehyde	LysY	No	Yes	No

	LysW-gamma-L-alpha-aminoadipate 6-semialdehyde -> LysW-gamma-L-lysine	LysJ	No	Yes	No
	LysW-gamma-L-lysine -> L-lysine	LysK	No	Yes	No
	2-oxoglutarate + acetyl-coA -> homocitrate	2.3.3.14	No	Yes	No
	Homocitrate <-> homo-cis-aconitate	4.2.1.114	No	Yes	No
	Homo-cis-aconitate <-> homoisocitrate	4.2.1.114 4.2.1.36	No	Yes	No
	Homoisocitrate <-> 2-oxoadipate	1.1.1.87 1.1.1.286	No	Yes	No
	2-oxoadipate <-> L-2-amino-adipate	2.6.1.39 2.6.1.57	No	Yes	No
Lysine Degradation	Saccharopine <-> L-2-aminoadipate 6-semialdehyde	1.5.1.9 1.5.1.10	No	Yes	No
	L-2-aminoadipate 6-semialdehyde -> L-2-aminoadipate	1.2.1.31	No	Yes	No
	L-2-aminoadipate <-> 2-oxoadipate	2.6.1.39	No	Yes	No
Alanine, Aspartate and Glutamate Metabolism	NH3 -> carbamoyl-phosphate	6.3.4.16	Yes	No	No
Glycine, Serine and Threonine Metabolism	Glycine -> guanidinoacetate	2.1.4.1	Yes	No	No
	Guanidinoacetate -> creatine	2.1.1.2	Yes	No	No
	Betaine -> dimethylglycine	2.1.1.5	Yes	No	No
	Dimethylglycine -> sarcosine	1.5.8.4	No	No	No
	Glycine <-> glyoxylate	2.6.1.44	Yes	Yes	No
	Glycine <-> threonine	4.1.2.5	Yes	Yes	Yes
	N-gamma-acetyl-L-2,4-diaminobutyrate -> L-ectoine	4.2.1.108	No	No	Yes
	L-ectoine -> 5-hydroxyectoine	1.14.11.- EctD	No	No	Yes
Alpha-Linolenic Acid Metabolism	Alpha-linolenic acid -> 13(S)-HpOTrE	1.13.11.12	No	Yes	No
	13(S)-HpOTrE -> 12,13-EOTrE	4.2.1.92	No	Yes	No
	12,13-EOTrE -> 12-OPDA	5.3.99.6	No	Yes	No
	12-OPDA -> OPC8	1.3.1.42	No	Yes	No
	OPC8 -> OPC8-CoA	OPCL1	No	Yes	No
	Trans-2-enoyl-OPC8-coA -> 3-oxo-OPC8-coA AND Trans-2-enoyl-OPC6-coA -> 3-oxo-OPC6-coA AND Trans-2-enoyl-OPC4-coA -> 3-oxo-OPC4-coA	MFP2	No	Yes	No
	JA-coA -> (+)-7-isojasmonate	3.1.2.-	No	Yes	No
Biosynthesis of Unsaturated Fatty Acids	Δ12, Δ15, Δ6, Δ5, Δ4		No	No	No
	Arrows 1, 3 and 4 going down		No	No	No
	Arrows 3 and 4 going up		No	No	No
	Δ9,12,15 -> alpha-linolenic acid	3.1.2.2	No	Yes	No

Synthesis and Degradation of Ketone Bodies	Acetoacetyl-coA + acetyl-coA -> (S)-3-hydroxy-3-methylglutaryl-coA	2.3.3.10	No	Yes	No
	Acetoacetate -> acetone	4.1.1.4	No	Yes	No
Glycerophospholipid Metabolism	CDP-diacyl-glycerol -> phosphatidyl-1D-myo-inositol	2.7.8.11	No	Yes	No
	Phosphatidyl-glycerophosphate -> phosphatidyl-glycerol	3.1.3.27	Yes	No	No
	Phosphatidyl-glycerol -> cardiolipin	2.7.8.-Cls	Yes	No	No
	1,2-diacyl-sn-glycerol 3-phosphate -> 1,2-diacyl-sn-glycerol	3.1.3.4	Yes	No	No
	Choline -> phosphocholine	2.7.1.32	No	Yes	No
	Phosphocholine -> CDP-choline	2.7.7.15	No	Yes	No
	CDP-choline -> phosphatidylcholine (lecithin)	2.7.8.2	No	Yes	No
	Acetaldehyde <-> ethanolamine	4.3.1.7	No	No	Yes
	Ethanolamine -> phosphoethanolamine	2.7.1.82	No	Yes	No
	Phosphoethanolamine -> CDP-ethanolamine	2.7.7.14	No	Yes	No
	CDP-ethanolamine -> phosphatidylethanolamine	2.7.8.1	No	Yes	No
Glycerolipid Metabolism	1,2-diacyl-sn-glycerol 3-phosphate -> 1,2-diacyl-sn-glycerol	3.1.3.4	Yes	Yes	No
Glyoxylate and Dicarboxylate Metabolism	Acetoacetyl-coA -> (R)-3-hydroxy-butanoyl-coA	1.1.1.36	No	No	No
	Crotonoyl-coA -> (2S)-ethyl-malonyl-coA	1.3.1.85	No	No	No
	Oxalate -> formate	4.1.1.2	No	No	Yes
	Hydroxypyruvate -> tartronate-semialdehyde	5.3.1.22	No	No	Yes
	H2O2 -> O2	1.11.1.6	No	No	Yes
	Glyoxylate -> glycine	2.6.1.45	Yes	No	No
	L-glutamate -> 2-oxoglutarate + glycine	GGAT	Yes	No	No
Glycolysis/ Gluconeogenesis	Alpha-D-glucose-1P -> alpha-D-glucose	3.1.3.10	Yes	No	No
	Pyruvate <-> L-lactate	1.1.1.27	Yes	No	No
	Glyceraldehyde-3P -> glycerate-3P (may be covered tho)	1.2.1.9	Yes	No	No
Pyruvate Metabolism	Phosphoenolpyruvate -> oxaloacetate	4.1.1.31	Yes	Yes	Yes
	Pyruvate <-> L-lactate	1.1.1.27	Yes	No	No
	Pyruvate <-> formate + acetyl-coA	2.3.1.54	Yes	Yes	No
Propanoate Metabolism	2-propyn-1-ol -> 2-propyn-1-al	1.1.2.8	No	No	Yes
	Malonate-semialdehyde -> propynoate	4.2.1.27	Yes	No	No
	Malonate-semialdehyde <-> 3-hydroxy-propanoate	1.1.1.59	Yes	Yes	No
	Lactate <-> lactoyl-coA	2.8.3.1	Yes	No	No
	Lactoyl-coA <-> acryloyl-coA	4.2.1.54	Yes	No	No

	2-oxo-butanoate <-> propanoyl-coA	2.3.1.54	Yes	No	No
	Propanoyl-coA <-> (S)-2-methyl-malonyl-coA	4.1.1.41	Yes	Yes	No
	2-methyl-cis-aconitate <-> (2S,3R)-3-hydroxybutane-1,2,3-tricarboxylate	4.2.1.99	Yes	No	No
	2-methyl-citrate <-> propanoyl-coA	2.3.3.5	No	No	No
Pentose Phosphate Pathway	D-glucose -> D-glucono-1,5-lactone	1.1.5.2	Yes	No	No
	D-glucono-1,5-lactone -> D-gluconate	3.1.1.17	Yes	No	Yes
	D-gluconate -> 6-phospho-D-gluconate	2.7.1.12	Yes	Yes	No
	2-dehydro-3-deoxy-D-gluconate -> 2-dehydro-3-deoxy-D-gluconate-6P	2.7.1.45	No	Yes	Yes
	2-dehydro-3-deoxy-D-gluconate-6P <-> pyruvate + D-glyceraldehyde-3P	4.1.2.14	No	Yes	Yes
	D-ribose-1,5P -> PRPP	2.7.4.23	Yes	No	No
Fructose and Mannose Metabolism	Alpha-D-glucose <-> D-sorbitol	1.1.1.21	No	Yes	No
	D-fructose-1P <-> D-fructose	2.7.1.69	No	No	Yes
	D-fructose -> beta-D-fructose-6P	2.7.1.4	No	Yes	Yes
	D-mannose -> D-mannose-6P	2.7.1.7	Yes	No	No
	GDP-D-rhamnose <-> GDP-4-oxo-6-deoxy-D-mannose	1.1.1.187	Yes	No	No
Pentose and Glucuronate Interconversions	Ribitol <-> D-ribulose	1.1.1.56	No	No	No
	L-arabinose <-> L-ribulose	5.3.1.4	No	No	Yes
	D-xylulose <-> D-xylose	5.3.1.5	No	No	Yes
	D-xylose <-> xylitol	1.1.1.21	No	No	No
	L-xylulose <-> L-xylulose-1P	2.7.1.5	No	Yes	Yes
Ascorbate and Aldarate Metabolism	UDP-D-glucuronate -> D-glucuronate-1P	2.7.7.44/ 2.7.7.64	No	Yes	No
	D-glucuronate-1P <-> D-glucuronate	2.7.1.43	No	Yes	No
	D-glucuronate <-> L-gulonate	1.1.1.19	No	Yes	No
	L-gulonate <-> L-gulono-1,4-lactone	3.1.1.17	No	Yes	Yes
	D-glucuronate <-> D-glucurono-lactone	3.1.1.19	No	Yes	No
	D-glucarate <-> 5-dehydro-4-deoxy-D-glucarate	4.2.1.40	No	No	Yes
Starch and Sucrose Metabolism	UDP-glucose -> sucrose	2.4.1.13	Yes	No	No
	Beta-D-fructose <-> beta-D-fructose-6P	2.7.1.4	No	No	Yes
	UDP-D-glucuronate -> UDP-D-xylose	4.1.1.35	Yes	No	No
	D-glucose <- Maltose	2.4.1.8	Yes	No	No
Amino Sugar and Nucleotide Sugar Metabolism	GlcNAc-6P <-> GlcNAc	2.7.1.59	Yes	No	No
	Man-6P <-> Man	2.7.1.7	Yes	No	No
	UDP-GlcA <-> GlcA-1P	2.7.7.44/	No	Yes	No

		2.7.7.64			
	GlcA-1P <-> GlcA	2.7.1.43	No	Yes	No
C5-Branched Dibasic Acid Metabolism	Propanoyl-coA + glyoxylate -> L-erythro-3-methylmalyl-coA	4.1.3.24	Yes	No	Yes
Butanoate Metabolism	Fumarate <-> maleate	5.2.1.1	No	Yes	Yes
	Maleate <-> (R)-malate	4.2.1.31	No	No	No
	(R)-2-acetoin <-> 2-acetolactate	4.1.1.5	No	No	No
	Pyruvate -> acetyl-coA	2.3.1.54	Yes	Yes	No
	Butanoyl-coA <-> crotonoyl-coA	1.3.8.1/ 1.3.1.44	No	No	No
Inositol Phosphate Metabolism	1D-myo-inositol <-> scyllo-inosose	1.1.1.18	No	No	Yes
	Scyllo-inosose <-> 3,5/4-trihydroxycyclohexa-1,2-dione	4.2.1.44	No	No	Yes
	5-deoxy-glucuronate -> 2-deoxy-5-keto-D-gluconate	5.3.1.-	No	No	Yes
	2-deoxy-5-keto-D-gluconate-6P -> malonic semialdehyde + dihydroxyacetone phosphate	4.1.2.29	No	No	No
Nitrogen Metabolism	Ammonia -> carbamoyl-P	6.3.4.16	Yes	No	No
	Carbamoyl-P -> carbamate	2.7.2.2	Yes	No	No
	Formamide -> formate + ammonia	3.5.1.49	No	No	Yes
Methane Metabolism	5-amino-6-ribitylaminouracil + 4-hydroxyphenylpyruvate -> 7,8-didemethyl-8-hydroxy-5-deazariboflavin	2.5.1.77	Yes	No	No
	Coenzyme F420 -> coenzyme F420H2	1.12.98.1	Yes	Yes	No
	5-methyl-THMPT + coenzyme M <-> methyl-coM + THMPT	2.1.1.86	No	Yes	No
	Methyl-coM -> methane	2.8.4.1	No	Yes	No
	Acetyl-coA -> pyruvate	1.2.7.1	Yes	No	No
	Phosphoenolpyruvate -> oxaloacetate	4.1.1.31	Yes	No	Yes
	L-malate -> malyl-coA	6.2.1.9	Yes	No	No
	Malyl-coA -> acetyl-coA + glyoxylate	4.1.3.24	Yes	No	No
	Glyoxylate + L-serine -> hydroxypyruvate + glycine	2.6.1.45	Yes	No	No
	Dihydroxyacetone -> dihydroxyacetone-phosphate	2.7.1.29	No	No	Yes
	CO2 -> formate	1.2.1.43	Yes	Yes	No
	CO2 + methanofuran <-> formyl-MFR	1.2.99.5	No	Yes	No
	Formyl-MFR + THMPT <-> methanofuran + N5-formyl-THMPT	2.3.1.101	No	Yes	No
	N5-formyl-THMPT <-> 5,10-methenyl-THMPT	3.5.4.27	No	Yes	No
	5,10-methenyl-THMPT + coenzyme F420 <-> 5,10-methylene-THMPT + coenzyme F420H2	1.5.99.9	Yes	Yes	No
	5,10-methylene-THMPT -> 5,10-methenyl-THMPT	1.5.1.15	Yes	Yes	No
Sulfur Metabolism	L-homoserine -> O-succinyl-L-homoserine (DUPLICATE)	2.3.1.46	Yes	Yes	No
Geraniol	Geraniol -> geranial	1.1.1.183	No	No	No

Degradation					
	Trans-geranyl-coA -> cis-geranyl-coA	5.2.1.-	No	No	No
	Citronellate -> citronellyl-coA	AtuH	No	No	No
	Citronellyl-coA -> cis-geranyl-coA	AtuD	No	No	No
	Cis-geranyl-coA -> isohexenylglutaconyl-coA	6.4.1.5	No	No	No
	Isohexenylglutaconyl-coA -> 3-hydroxy-3-isohexenylglutaryl-coA	4.2.1.57 AtuE	No	No	No
Limonene and Pinene Degradation	Alpha-pinene oxide -> cis-2-methyl-5-isopropylhexa-2,5-dienal	5.5.1.10	No	No	No
	S-limonene -> perillyl alcohol	1.14.13.49	No	No	No
	3-hydroxy-2,6-dimethyl-5-methylene-heptanoyl-coA -> 2,6-dimethyl-5-methylene-3-oxo-heptanoyl-coA	1.1.-.-	No	No	No
	3-isopropylbut-3-enoyl-coA -> 3-isopropylbut-3-enoic acid	3.1.2.-	No	No	No
	R-limonene -> trans-carveol	1.14.13.80	No	No	No
	4S-carvone -> 1R,4S-isodihydro-carvone	1.3.99.25	No	No	No
	1R,4S-isodihydro-carvone -> 4S,7R-4-isopropenyl-7-methyl-2-oxo-oxepanone	1.14.13.105	No	No	No
	4S,7R-4-isopropenyl-7-methyl-2-oxo-oxepanone -> 3S-6-hydroxy-3-isopropenyl-heptanoate	3.1.1.83	No	No	No
	2-hydroxy-4-isopropenyl-cyclohexane-1-carboxyl-coA -> 4-isopropenyl-2-ketocyclohexane-1-carboxyl-coA	1.1.-.-	No	No	No
	4-isopropenyl-2-ketocyclohexane-1-carboxyl-coA -> 3-isopropenyl-pimelyl-coA	3.7.1.-	No	No	No
Porphyrin and Chlorophyll Metabolism	Co-precorrin 5B -> co-precorrin 6A	2.1.1.195	Yes	No	No
	Co-precorrin 7 -> co-precorrin 8X	2.1.1.196	Yes	No	No
	Cob(II)yrinate a,c diamide -> cob(I)yrinate a,c diamide	1.16.8.1	Yes	No	No
Ubiquinone and Other Terpenoid-Quinone Biosynthesis	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate -> (1R,6R)-2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate	4.2.99.20	Yes	No	No
	1,4-dihydroxy-2-naphthoyl-coA -> 1,4-dihydroxy-2-naphthoate	3.1.2.28	Yes	No	No
	Menaquinone -> menaquinol AND Phylloquinone -> phylloquinol	1.6.5.2	Yes	No	Yes
Thiamine Metabolism	[ThiI]-SSH -> [ThiS]-COSH	ThiI	No	No	No
	[ThiS]-COSH -> [ThiS]-COSS-[ThiF]	2.7.7.73	No	No	No
Riboflavin Metabolism	5-amino-6-(5-phospho-D-ribitylamino) uracil -> 5-amino-6-ribityl-aminouracil	3.1.3.-	No	No	No

Vitamin B6 Metabolism	3-amino-2-oxopropyl phosphate -> Pyridoxine phosphate	2.6.99.2	Yes	No	No
	O-phospho-4-hydroxy-L-threonine -> (2S)-2-amino-3-oxo-4-phosphono butyanoate	1.1.1.262	Yes	No	No
	D-erythrose 4-phosphate <-> 4-phospho-D-erythronate	1.2.1.72	Yes	No	No
	4-phospho-D-erythronate <-> 2-oxo-3-hydroxy-4-phosphobutanoate	1.1.1.290	Yes	No	No
Nicotinate and Nicotinamide Metabolism	N-ribosyl-nicotinamide -> nicotinamide D-ribonucleotide	2.7.1.22	No	No	No
	Nicotinate D-ribonucleoside -> nicotinate D-ribonucleotide	2.7.1.173	No	No	No
Pantothenate and CoA Biosynthesis	N-pantothenoyl-cysteine -> Pantetheine	4.1.1.30	Yes	No	No
	Pantetheine -> (R)-pantothenate	3.5.1.92	Yes	No	No
	Acyl-carrier protein -> 4'-phosphopantetheine	3.1.4.14	Yes	No	No
	Uracil <-> 5,6-dihydrouracil	1.3.1.1 1.3.1.2	No	No	No
	N-carbamoyl-beta-alanine -> beta-alanine	3.5.1.6	No	No	No
Biotin Metabolism	3-hydroxy-pimeloyl-[acp] methyl ester -> enoyl-pimeloyl-[acp] methyl ester	FabZ	No	Yes	No
	Pimeloyl-[acp] methyl ester -> pimeloyl-[acp]	BioH BioG	No	Yes	No
	Long-chain-acyl-[acp] -> pimeloyl-[acp]	BioI	Yes	Yes	No
	Pimelate -> pimeloyl-coA	6.2.1.14 BioW	Yes	Yes	No
	N6-D-biotinyl-L-lysine (biocytin) -> L-lysine + biotin	3.5.1.12	Yes	Yes	No
One Carbon Pool by Folate	5,6,7,8-tetrahydrofolate (THF) -> 10-formyl-THF	6.3.4.3	Yes	Yes	No
	5,10-methenyl-THF <-> 5,10-methylene-THF	1.5.1.15	Yes	Yes	No
	5,10-methylene-THF -> 5-methyl-THF	1.5.1.20	Yes	Yes	Yes
Folate Biosynthesis	7,8-dihydroneopterin 3'-triphosphate -> dihydroneopterin	3.1.3.1	Yes	Yes	Yes
Peptidoglycan Biosynthesis	UDP-MurNAc-L-Ala-D-Glu -> UDP-MurNAc-L-Ala-gamma-D-Glu-L-Lys	6.3.2.7	No	No	No
	Und-PP-MurNAc-(GlcNAc)-L-Ala-gamma-D-Glu-L-Lys-D-Ala-D-Ala -> Und-PP-MurNAc-(GlcNAc)-L-Ala-gamma-D-Glu-L-Lys-(L-Ala)-D-Ala-D-Ala	murM	No	No	No
	Und-PP-MurNAc-(GlcNAc)-L-Ala-gamma-D-Glu-L-Lys-(L-Ala)-D-Ala-D-Ala -> Und-PP-MurNAc-(GlcNAc)-L-Ala-gamma-D-Glu-L-Lys-(L-Ala)2-D-Ala-D-Ala	murN	No	No	No
Lipopolysaccharide Biosynthesis	D-glycero-alpha-D-manno-heptose-1-P -> GDP-D-glycero-alpha-D-manno-heptose	2.7.7.71	No	No	No

	Lauroyl-KDO2-lipid IV(A) + myristoyl-ACP -> KDO2-lipid(A)	MsbB	No	No	No
Taurine and Hypotaurine Metabolism	Taurine + pyruvate <-> sulfoacetaldehyde + L-alanine	2.6.1.77	Yes	No	No
	Sulfoacetaldehyde -> acetyl phosphate + sulfite	2.3.3.15	No	No	No
	Sulfoacetaldehyde -> isethionate	1.1.1.313	No	No	No
Beta-Alanine Metabolism	Spermidine -> 1,3-diamino-propane + 4-aminobutanal	1.5.99.6	No	No	No
	Beta-alanine <-> N-carbamoyl-beta-alanine	3.5.1.6	No	No	No
	5,6-dihydrouracil <-> uracil	1.3.1.1 1.3.1.2	No	No	No
	3-hydroxy-propanoyl-coA <-> 3-hydroxy-propanoate	3.1.2.4	Yes	No	No
	3-hydroxy-propanoate <-> malonate semialdehyde	1.1.1.59	No	No	No
	Malonyl-coA -> Acetyl-coA	4.1.1.9	No	No	No
Selenocompound Metabolism	Selenocysteine -> Se-methyl-seleno-cysteine	2.1.1.-	No	No	No
D-Glutamine and D-Glutamate Metabolism	L-glutamate <-> 2-oxoglutarate	1.4.1.3	Yes	No	No
D-Arginine and D-Ornithine Metabolism	D-arginine -> D-ornithine	3.5.3.10	Yes	No	No
	D-arginine <-> L-arginine AND D-ornithine <-> L-ornithine	5.1.1.9	No	No	No
Cyanoamino Acid Metabolism	Cyanide -> L-3-cyanoalanine	4.4.1.9	No	No	No
Glutathione Metabolism	Glutathione(GSH) + RX -> R-S-glutathione	2.5.1.18	No	No	No
	L-gamma-glutamylcysteine + glycine -> glutathione (GSH)	6.3.2.3	Yes	No	No
	Glutathione (GSH) -> glutathione disulfide (GSSG)	1.11.1.9	No	Yes	Yes
Streptomycin Biosynthesis	dDTP-L-dihydro-streptose -> O-1,4-alpha-L-dihydro-streptosyl-streptidine-6P	2.4.2.27	No	No	No
	Streptomycin-6P -> streptomycin	3.1.3.39	No	No	No
	Streptomycin -> streptomycin-6P	2.7.1.72	No	No	No
Benzoate Degradation	4-hydroxy-benzoyl-coA -> 4-hyroxy-benzoate	3.1.2.23	No	No	In Rjost
	3,4-dihydroxy-benzoate -> beta-carboxy-muconate	1.13.11.3	No	No	Yes
	3-oxoadipate -> 3-oxoadipyl-coA	2.8.3.6	Yes	Yes	No
	2-hydroxy-muconate -> gamma-oxalocrotonate	5.3.2.-	No	No	In Rjost
	Glutaryl-coA -> glutaconyl-coA	1.3.99.32	Yes	Yes	No
	Glutaconyl-coA -> crotonoyl-coA	4.1.1.70	Yes	Yes	No
Chlorocyclohexane and Chlorobenzene	3,4,6-trichloro-cis-1,2-dihydroxy-cyclohexa-3,5-diene -> 3,4,6-trichloro-catechol	1.3.1.19	No	No	No

Degradation	AND Cis-dihydrobenzenediol -> catechol AND 3-chloro-cis-1,2-dihydroxy-cyclohexa-3,5-diene -> 3-chlorocatechol				
	2-chloro-cis,cis-muconate -> trans-4-carboxymethylene-but-2-en-4-olide AND 3-chloro-cis,cis-muconate -> 2-chloro-5-oxo-2,5-dihydrofuran-2-acetate AND 2,3,5-trichloro-cis,cis-muconate -> 2,5-dichloro-carboxymethylene-but-2-en-4-olide AND Tetrachloro-cis,cis-muconate -> 2,3,5-trichloro-dienelactone	5.5.1.7	No	No	No
Xylene Degradation	2-hydroxy-5-methyl-cis,cis-muconate -> 2-oxo-5-methyl-cis-muconate	5.3.2.-	No	No	Yes
	p-xylene -> 4-methylbenzyl-alcohol AND o-xylene -> 2-methylbenzyl-alcohol AND m-xylene -> 3-methylbenzyl-alcohol	XylM XylA	No	No	No
Nitrotoluene Degradation	4-hydroxylamino-2,6-dinitrotoluene -> 2,4-diamino-6-nitrotoluene AND 2,hydroxylamino-4,6-dinitrotoluene -> 2,4-diamino-6-nitrotoluene	1.7.1.-	No	No	No
	2,4-diamino-6-hydroxylaminotoluene -> 2,4,6-triamino-toluene	1.8.99.3	No	No	No
Styrene Degradation	Phenylacetonitrile -> phenylacetamide AND Acrylonitrile -> acrylamide	4.2.1.84	No	Yes	Yes
	Acrylonitrile -> acrylate	3.5.5.7	No	Yes	Yes
	Acrylyl-coA -> lactoyl-coA	4.2.1.54	Yes	Yes	No
	Lactoyl-coA L-lactate	2.8.3.1	Yes	Yes	No
Atrazine Degradation	Cyanuric acid -> biuret	3.5.2.15	No	No	Yes
	Biuret -> allophanate	3.5.1.84	No	No	No
	Hydroxyatrazine -> N-isopropylammelide	3.5.99.3	No	No	Yes
	N-isopropylammelide -> cyanuric acid	3.5.99.4	No	No	No
Naphthalene Degradation	1-methylnaphthalene -> cis-1,2-dihydroxy-1,2-dihydro-8-methylnaphthalene AND 2-hydroxymethyl-naphthalene -> cis-1,2-dihydroxy-1,2-dihydro-7-hydroxy-methylnaphthalene	1.14.12.12	No	No	No
	cis-1,2-dihydroxy-1,2-dihydro-8-methylnaphthalene -> 1,2-dihydroxy-8-methylnaphthalene AND cis-1,2-dihydroxy-1,2-dihydro-7-hydroxy-methylnaphthalene -> 1,2-	1.3.1.29	No	No	No

	dihydroxy-7-hydroxymethylnaphthalene				
	1,2-dihydroxy-8-methylnaphthalene -> 2-hydroxy-8-methyl-chromene-2-carboxylate AND 1,2-dihydroxy-7-hydroxymethylnaphthalene -> 2-hydroxy-7-hydroxymethyl-chromene-2-carboxylate	1.13.-.-	No	No	No
	2-hydroxy-8-methyl-chromene-2-carboxylate -> 2-hydroxy-3-methyl-benzalpyruvate AND 2-hydroxy-7-hydroxymethyl-chromene-2-carboxylate -> 2-hydroxy-4-hydroxymethyl-benzalpyruvate	5.3.99.-	No	No	No
Aminobenzoate Degradation	Benzonitrile -> benzamide	4.2.1.84	Yes	Yes	Yes
	4-nitrophenol -> 4-nitrocatechol	1.14.13.29	No	No	No
	Cyclopropane-carboxyl-coA -> crotonoyl-coA	5.5.1.-	No	No	No
Chloroalkane and Chloroalkene Degradation	2-chloroethanol -> chloroacetaldehyde	1.1.2.8	Yes	Yes	Yes
	Trans-3-chloroacrylic acid -> malonate semialdehyde	CaaD	No	No	No
	Cis-3-chloroacrylic acid -> malonate semialdehyde	Cis-CaaD	No	No	No
	Malonate semialdehyde -> acetaldehyde	4.1.1.-	No	No	No
Toluene Degradation	3-methyl-muconolactone -> 4-methyl-3-oxoadipate-enol-lactone	5.5.1.-	No	No	No
	4-methyl-3-oxoadipate-enol-lactone -> 4-methyl-3-oxoadipate	3.1.1.-	No	No	No
	Toluene -> 3-hydroxytoluene AND Toluene -> benzyl alcohol AND Toluene -> 4-hydroxytoluene	1.14.13.-	No	No	No
	Benzaldehyde -> benzoate AND 4-hydroxy-benzaldehyde -> 4-hydroxy-benzoate	1.2.1.28	No	No	Yes
	4,6-dichloro-3-methyl-cis-1,2-dihydroxycyclohexa-3,5-diene -> 4,6-dichloro-3-methylcatechol	1.3.1.-	No	No	No
	4,6-dichloro-3-methylcatechol -> 3,5-dichloro-2-methyl-muconate	1.13.11.-	No	No	No
	3,5-dichloro-2-methyl-muconate -> 2-chloro-5-methyl-cis-dienelactone	5.5.1.7	No	No	No
Ethylbenzene Degradation	Cis-1,2-dihydroxy-2,3-dihydroethylbenzene -> 2,3-dihydroxy-ethylbenzene	1.3.1.66	No	No	No

Caprolactam Degradation	Cyclohexane -> cyclohexanol	1.14.15.-	No	Yes	No
	6-hexanolide -> 6-hydroxyhexanoate	3.1.1.17	Yes	Yes	Yes
Bisphenol Degradation	4-hydroxyphenyl acetate -> hydroquinone	3.1.1.2	No	No	No
	Bis(4-hydroxyphenyl)-methanol -> 4,4'-dihydroxy-benzophenone	1.1.-.-	Yes	No	No
Polycyclic Aromatic Hydrocarbon Degradation	Cis-3,4-dihydroxy-3,4-dihydrophenanthrene -> 3,4-dihydroxy-phenanthrene	1.3.1.49	No	No	No
	2-hydroxy-2H-benzo[h]chromene-2-carboxylate -> is-4-(1'-hydroxy-naphth-2'-yl)-2-oxobut-3-enoate	5.1.2.-	No	No	No
	1-hydroxy-2-naphthoate -> cis-2'-carboxy-benzalpyruvate	1.13.11.38	No	No	Yes
	2-carboxy-benzaldehyde -> phthalate	1.2.1.78	No	No	No
	Phthalate-4,5-cis-dihydrodiol -> 4,5-dihydroxy-phthalate	1.3.1.64	No	No	No
	4,5-dihydroxy-phthalate -> 3,4-dihydroxy-benzoate	4.1.1.55	No	No	Yes
	1-methoxy-pyrene -> 1-methoxypyrene-6,7-oxide	1.14.-.-	No	No	Yes